Recent progress in massively parallel sequencing platforms has enabled genome-wide characterization of DNA-associated proteins using the combination of chromatin immunoprecipitation and sequencing (ChIP-seq). Although a variety of methods exist for analysis of the established alternative ChIP microarray (ChIP-chip), few approaches have been described for processing ChIP-seq data. To fill this gap, we propose an analysis pipeline specifically designed to detect protein-binding positions with high accuracy. Using previously reported data sets for three transcription factors, we illustrate methods for improving tag alignment and correcting for background signals. We compare the sensitivity and spatial precision of three peak detection algorithms with published methods, demonstrating gains in spatial precision when an asymmetric distribution of tags on positive and negative strands is considered. We also analyze the relationship between the depth of sequencing and characteristics of the detected binding positions, and provide a method for estimating the sequencing depth necessary for a desired coverage of protein binding sites.

The combination of chromatin immunoprecipitation and microarray hybridization (ChIP-chip) has been used extensively to determine chromosome binding patterns of DNA-associated proteins. Nonetheless, several recent studies have demonstrated that newly developed high-throughput sequencing methods provide marked improvements over microarray measurements. Although sequencing techniques were previously combined with both chromatin immunoprecipitation and sequence tagging methods, the new generation of high-throughput sequencing platforms provides orders-of-magnitude increases in the numbers of sequences generated. This permits cost-effective genome-wide mapping of the binding sites of many proteins of interest, such as transcription factors, insulators or chromatin modifying enzymes.

Processing of ChIP-chip data has focused on compensating for array limitations, such as probe-specific behavior, dye bias and tiling resolution. Although the ChIP-seq approach avoids such biases and can provide greater sensitivity and specificity while requiring a much smaller amount of starting material, ChIP-seq data pose their own unique challenges. Given that the rate of sequencing errors varies between and within the sequenced reads, what range of sequence tag quality should be tolerated when aligning tags to the genome? What background tag distribution is appropriate for assessing the significance of observed binding positions? What is the required depth of sequencing? And how can this information be used to accurately determine protein binding positions?

Here we describe a data processing pipeline optimized for detection of localized protein binding positions from unpaired sequence reads. We illustrate the proposed pipeline on published data sets produced using the Solexa platform for genome-wide binding of the transcription factors NRSF (neuron-restrictive silencer factor), CTCF (CCCTC binding factor) and STAT1 (signal transducers and activators of transcription). The alignment procedure is enhanced to maximize the number of informative tags, based on the strand-specific pattern of tag distribution expected around a binding position. Filtering and background correction steps reduce the numbers of false-positive determinations. We compare the performance of our three algorithms and two previously described computational methods for calling specific binding positions, and show that some methods provide greater specificity and more precise estimates of positions. The final step of the proposed pipeline examines the saturation level of detected binding positions to estimate how much additional sequencing may be necessary.

**RESULTS**

**Tag distribution around protein binding positions**

In general, immunoprecipitation selects a set of overlapping DNA fragments around bound positions. High-throughput sequencing identifies short (~35 bp for the Solexa or SOLiD platforms) tags on the 5' ends of fragments from either DNA strand. The positions of the tags are then determined by aligning them to the genome assembly, with ambiguous alignments typically discarded. The resulting spatial distribution of tag occurrences around a stable binding position will therefore show separate peaks of tag density on positive and negative strands. The distance between the peaks should reflect the size of the protected region, although it may also be influenced by the size distribution of the DNA fragments. This distance does not exhibit strong dependency on the number of tags within the peaks (Supplementary Table 1 online).

A genome-wide signature of this tag pattern can be assessed by calculating the cross-correlation of positive- and negative-strand tag
 densities, shifting the strands relative to each other by increasing distance. All of the examined data sets exhibit a clear peak in the strand cross-correlation profile, corresponding to the predominant size of the protected region (Fig. 1d and Supplementary Fig. 1 online). The magnitude of the peak reflects the fraction of tags in the data set that appears in accordance with the expected binding tag pattern. In an ideal case, when all of the sequenced tags participate in such binding patterns, the correlation magnitude reaches a maximum value. Conversely, the magnitude decreases as tag positions are randomized (Supplementary Fig. 2 online).

Using variable-quality tag alignments

Although some tags align perfectly with the reference genome, others align only partially, with gaps or mismatches. Poorly aligned tags may result from experimental problems such as sample contamination, correspond to polymorphic or unassembled regions of the genome, or reflect sequencing errors. For the Solexa platform, the sequencing errors are more abundant toward the 3′ ends of the sequenced fragments, frequently resulting in partial alignments that include only the portions of the tags near the 5′ ends. We estimate that this increase in mismatch frequencies toward 3′ termini accounts for 41–75% of all observed mismatches in the examined data sets (Supplementary Fig. 3 online). As it is not unusual to have >50% of the total tags result in only partial alignment, inclusion of tags that are partially aligned but still informative is important for optimizing use of any data set\cite{11,12}. We therefore chose to use the length of the match and the number of nucleotides covered by mismatches and gaps to classify the quality of tag alignment (Table 1 and Supplementary Table 2 online).

Given a classification of tags by quality of alignment, we propose to use the strand cross-correlation profile to determine whether a particular class of tags should be included in further analysis. A set of tags informative about the binding positions should increase cross-correlation magnitude, whereas a randomly mapped set of tags should decrease it (Supplementary Fig. 2). Using this approach for the NRSF data set (Fig. 2), we found that alignments with matches spanning at least 18 bp and zero mismatches improved the cross-correlation profile. However, only full-length (25 bp) matches should be considered for tags with two mismatches. Using this criterion to accept tags increased their number over the set of perfectly aligned tags by 27% for the NRSF data set, 30% for the CTCF data set and 36% for the STAT1 data set (Supplementary Fig. 4 online). The incorporation of these tags improved sensitivity and accuracy of the identified binding positions (Supplementary Fig. 5 online).

Controlling for background tag distribution

The statistical significance of the tag clustering observed for a putative protein binding position depends on the expected background pattern. The simplest model assumes that the background tag density is distributed uniformly along the genome and independently between the strands\cite{11}. In addition to the NRSF ChIP sample, Johnson et al.\cite{2} have sequenced a control input sample, providing an experimental assessment of the background tag distribution. We found that the background tag distribution exhibits a degree of clustering that is

### Table 1 Classification of tag alignments based on the length of the match and the number of mismatches

| Match Length | Mismatches | Count |
|--------------|------------|-------|
| 16           | 0          | 63,388 |
| 17           | 0          | 50,613 |
| 18           | 0          | 34,707 |
| 19           | 0          | 21,230 |
| 20           | 0          | 16,775 |
| 21           | 0          | 14,453 |
| 22           | 0          | 11,068 |
| 23           | 0          | 6,556  |
| 24           | 0          | 54,455 |
| 25           | 0          | 1,234,829 |
| 16           | 1          | 16,625 |
| 17           | 1          | 25,991 |
| 18           | 1          | 24,715 |
| 19           | 1          | 23,431 |
| 20           | 1          | 17,540 |
| 21           | 1          | 12,705 |
| 22           | 1          | 11,068 |
| 23           | 1          | 6,556  |
| 24           | 1          | 54,455 |
| 25           | 1          | 1,234,829 |

The table gives the number of NRSF data set tags whose best alignment falls within each class, as defined by the length of alignment (columns) and the number of mismatches (rows). The tags from the NRSF data set were aligned using BLAT. The number of mismatches includes the number of nucleotides covered by gaps.
Our methods take advantage of the strand-specific tag pattern observed at binding positions (Fig. 1c). The first such method, window tag density (WTD), is similar to XSET but scores positions based on the strand-specific tag counts upstream and downstream of the examined position (Fig. 4a). The second method, matching strand peaks (MSP), determines local peaks of strand-specific tag density and identifies positions surrounded by positive- and negative-strand peaks of a comparable magnitude at the expected distance (Fig. 4b). The third method, mirror tag correlation (MTC), scans the genome to identify positions exhibiting pronounced positive- and negative-strand tag patterns that mirror each other (Fig. 4c). The source code is available online (Supplementary Source Code), and an up-to-date R package can be downloaded at http://compbio.med.harvard.edu/Supplements/ChIP-seq.

Although a complete list of true binding sites is not known for any of the examined data sets, all three proteins exhibit known binding sequence specificities. While the binding detection methods described in this work do not rely on sequence information, we used high-scoring sequence motif instances to assess relative performances of different binding detection methods. In doing so, we assume only that the high-scoring motif instances contain a representative subset of true binding positions, and do not require all high-scoring motifs to be bound, or that all true binding sites exhibit a motif signature. We evaluated performance using canonical sequence motifs for binding by NRSF and CTCF\(^{14,15}\), and the gamma-activated site (GAS) motif as a predictor of STAT1 binding\(^{5,11}\). The binding detection methods provide peak magnitude scores associated with the identified binding positions, thus allowing prioritization of binding positions determined by each method.
To compare the sensitivity of different methods, we selected increasing numbers of top binding positions returned by each method and examined the fraction of motif occurrences for which a binding position was identified (Fig. 4d). We found that 89% of the selected highest-scoring NRSF motif matches coincided with the detected binding positions. The motif coverage rate clearly exceeds that expected from random prediction, enabling comparison of the relative performances of the different binding detection methods. Except for MSP and CSP, all of the methods achieve similarly high motif coverage. The CSP method performs worse for the more prominent binding positions (top 500), whereas the MSP approach performs poorly throughout the entire range. Analyses of STAT1 and CTCF binding show analogous results in terms of the relative performances of the different methods (Supplementary Fig. 7 online).

These results are also confirmed by analysis of PCR-validated binding loci from the literature (11,15) (Supplementary Figs. 8 and 9 online). We note that the motif and PCR-validated test sets represent only a fraction of true binding sites. As this fraction is smaller for CTCF and STAT1, larger sets of top binding positions are used to illustrate test-set coverage by different methods.

The background subtraction methods outlined in the previous section improve the NRSF motif coverage, reaching the same level of coverage at up to 11% fewer top binding positions (Supplementary Fig. 10 online). The corrections have little effect on the top 1,500 binding positions, which are associated with higher tag counts than any false-positive peaks arising from uneven background. The background-driven false-positive positions are generally smaller in magnitude and begin to influence predictions as more binding positions are considered.

**Precision of binding positions**

To evaluate the spatial precision with which protein-binding positions are identified by different methods, we have analyzed the distances between predicted positions and locations of high-scoring motif hits (Fig. 5a). For the NRSF data set, the WTD method predicts binding positions with the greatest precision, with >60% of predicted peaks located within 10 bp of the motif center (Fig. 5b and Supplementary Fig. 11a online). It is followed by the XSET, MTC and MSP methods, with CSP calling ~40% of peaks within 10 bp of the motifs. Background corrections have limited effect on the precision of the predicted positions, with only the WTD method showing 3% improvement for strong binding positions (data not shown).

For the CTCF and STAT1 predictions, however, the MTC method achieves better precision than WTD (Fig. 5c,d and Supplementary Fig. 11b,c). The difference can be explained by the properties of the tag distribution immediately near the center of the protected region. Unlike WTD and XSET, the MTC method does not take into account tags within the central region (30 bp) when scoring binding positions. Altering the MTC method to take such positions into account reduces the precision of the determined binding positions to a level similar to the WTD predictions. Examining the overall distribution of tag positions relative to high-scoring motif hits, we found that CTCF and STAT1 showed unexpected peaks of tag density immediately adjacent (within 10–15 bp) to the motif position (Supplementary Fig. 12 online). This pattern, in which small sets of negative strand tags appear immediately upstream of the protected region and are mirrored by the positive strand tags immediately downstream, may result from cross-linking interactions occurring beyond the central protected region (Fig. 1b, broken line). As a result, peak detection methods that take into account the tags near the central region tend to call positions 15–20 bp upstream or downstream of the true binding site.

**Statistically significant positions**

The binding detection methods should limit the resulting binding positions to those that are not likely to have occurred by chance. The desired level of statistical significance is commonly given in terms of a false discovery rate (FDR) or the number of expected false-positive positions (E-value).
The detection methods can then use background tag distribution to determine the minimal binding position score satisfying the specified level of significance. Many false-positive calls originate from the large anomalous regions described earlier. These systematic errors can be filtered before determination of significance thresholds. Based on the input sample data for the NRSF, we found a total of 2,755 binding positions for the FDR threshold of 0.01 using the WTD method. This closely corresponds to the number of top peaks that was required to achieve maximal coverage of high-scoring motif positions used in the previous sections (Fig. 4d).

In the absence of an empirical estimate of the background tag distribution, it may be possible to rely on an analytical model. The simplest such model is a spatial Poisson process where the tags are uniformly distributed across the accessible regions of the genome. However, because the true background tag distributions exhibit a significant degree of clustering properties.

A closer estimate of statistical thresholds may be obtained by using randomization models with different bin sizes. For the FDR of 0.01, a randomization model that maintains together tags occurring at exactly the same position in the genome results in a comparable number of NRSF-binding positions (2,985). We used such randomization to determine the number of statistically significant binding positions for the CTCF (2,398 positions for an FDR of 0.01) and STAT1 (44,921 positions for an FDR of 0.01) data sets. Matching the number of binding positions for more stringent FDR values requires larger tag randomization blocks (Supplementary Table 3), indicating that simple randomization strategies cannot properly account for the background clustering properties.

Testing for sufficient sequencing depth

To assess whether the sequencing depth has reached a saturation point beyond which no additional binding sites are detected, we analyzed how the set of the predicted binding sites changed when only a subset of the tag data was used for prediction. Sampling increasing fractions of the tag data, we determined binding positions and compared these predictions with the set of reference binding sites identified from the complete data (Fig. 6a and Supplementary Fig. 13 online).

If the sequencing depth has moved beyond the saturation point, it would be possible to arrive at the reference set using only a subset of the tag data. We found, however, that none of the three data sets reached such a saturation point (horizontal asymptote), and that the fraction of the concordant binding positions decreased when even a small fraction of the tag data were omitted. This indicates that additional binding sites are being continuously identified with increasing sequencing depth. The observed trend holds for a range of FDR thresholds (Supplementary Fig. 13): although the slope of
the saturation curve can be reduced by setting a considerably more stringent FDR threshold, this results in a significantly smaller number of binding sites.

To understand the properties of the binding site coverage, we examined tag counts associated with high-scoring sequence motifs (Fig. 6b and Supplementary Fig. 14 online). In all three data sets, the distribution of tag counts showed a very wide dynamic range. Whereas some positions had hundreds of tags, others barely rose above the expected background counts. Moreover, these distributions appeared to be continuous in that they did not show distinct subpopulations of binding positions. This suggests that increasing sequencing depth may allow a greater number of weak binding positions to be distinguished without a qualitative threshold that would define a complete set of binding sites.

As more pronounced binding positions are identified using smaller sequencing depth, an experiment of given depth may saturate detection of the binding positions that exceed a certain tag enrichment ratio relative to the background. We refer to this enrichment ratio as the minimal saturated enrichment ratio (MSER). The saturation criteria that define the maximal acceptable slope of the saturation curve (Fig. 6a) can be formulated as a requirement for stability of the set of predicted binding sites. For instance, we require 99% agreement in the set of binding positions when the data set is reduced by 10^7 tags. Using NRSF input tag data to determine the confidence intervals for the enrichment ratio of each binding position, we found that the achieved sequencing depth was sufficient to saturate detection of binding positions with tag enrichment ratios significantly above 7.5 (P-value < 0.05; Fig. 6a and Supplementary Fig. 15 online). Of the 2,755 NRSF binding positions detected at an FDR of 0.01, 1,879 (68%) had enrichment ratios significantly greater than the MSER value of 7.5 (Supplementary Fig. 13). We note that a particular MSER value does not imply that all of the true binding positions of that fold-enrichment have been discovered; instead, it indicates that new binding positions with enrichment significantly higher than the MSER value are being detected at a sufficiently slow rate. A potential range of true enrichment ratios can be assessed from the enrichment confidence intervals calculated for each binding position (Supplementary Fig. 16 online). As estimation of the enrichment ratio confidence intervals also depends on the amount of information available about the background tag distribution, input data sets of similar genomic coverage should be used when comparing different MSER values.

For practical purposes, it is important to be able to predict the number of tags required to saturate detection of peaks above a given target enrichment ratio. The relationship between the number of tags and the MSER settles into a dependency that can be extrapolated using a log-log model (Fig. 6c). We predict, for instance, that 1.2 \times 10^6 more sequence tags would be required to saturate detection of binding positions with an enrichment ratio twofold higher (P-value < 0.05). The MSER values and extrapolations depend on the saturation criteria and on methods used to calculate enrichment confidence intervals (Supplementary Fig. 17 online).

Increasing the sequencing depth is also likely to lead to increased accuracy of the determined binding positions. Using the NRSF data set, we analyzed how the mean distance between the detected binding positions and sequence motifs depends on the number of tags used for predictions. Our results show that accuracy indeed improved with the increasing number of tags (Supplementary Fig. 18 online). The improvement, however, was minor: the accuracy decreased by only several base-pairs even when the number of tags was halved.

**DISCUSSION**

Analysis of protein-DNA interactions using high-throughput short sequencing poses a number of novel computational challenges. We show that many aspects of the processing pipeline can be specifically tailored to improve detection of binding positions.

The protein-binding positions exhibit a strand-specific pattern of tag occurrences. We show that a genome-wide signature of such a pattern can be obtained, with strand cross-correlation of tag density providing a quick assessment of data set quality and binding characteristics. The proposed alignment procedure also relies on this signature to determine the range of alignment quality that is informative about the binding positions. In our implementation, we have...
used a simple classification of tag alignment quality, based on the number of nucleotide mismatches. The same procedure can be applied to more elaborate measures of alignment quality, such as those incorporating confidence in specific base calls.

The examination of the input sequencing clearly indicates that experimental assessment of the background tag distribution is necessary for accurate evaluation of ChIP-seq data. The background distribution is far from uniform and, in some cases, shows tag-density patterns similar to those expected from true binding positions. We demonstrate that the knowledge of such distribution is instrumental for accurately assessing and reducing rates of false-positive predictions. As additional data sets become available, it will be important to analyze the degree to which tag profiles of input or no-antibody measurements differ between independent experiments.

Comparison of different binding prediction algorithms shows that even though several methods can reach optimal sensitivity, there is a considerable variation in the accuracy of the identified binding positions. Although the MTC method provided more accurate positions for CTCF and STAT1 binding, the WTD method was better at identifying precise positions of NRSF binding. The difference can be attributed to the consideration of tag patterns immediately near the center of the binding pattern, which show qualitative differences between the NRSF data set and the CTCF or STAT1 data sets (Supplementary Fig. 12). As the NRSF binding tag pattern was more consistent with the basic expectations, we recommend using the WTD method when the tag pattern cannot be examined beforehand on a set of expected binding positions. It remains to be seen, however, which tag pattern will be typical of other experiments and whether both patterns can be efficiently handled by a single method.

The ability to evaluate and predict the sequencing depth requirement is an important aspect of ChIP-seq studies. Our analyses demonstrate that none of the three examined data sets definitively reached a point of saturation at which the set of determined binding positions stabilized. The binding positions exhibited a very wide range of enrichment ratios so that additional sequencing revealed an increasing number of weaker binding sites. This bears some resemblance to other genomics studies. In genome-wide association studies, for instance, increasing the sample size allows one to find more and more loci with smaller lod scores; in gene expression studies, it leads one to find more and more genes with a statistically significant but smaller fold-change. In practical terms, this lack of saturation point has profound implications in study design. It suggests that it would be difficult to define a ‘sufficient’ depth of sequencing and that other criteria must be specified.

We therefore propose that the sequencing depth requirements should be evaluated with respect to a specific target enrichment ratio of the binding positions. Toward that end, we provide a method to determine the minimal fold-enrichment ratio above which the detection of binding positions has been saturated (that is, stabilized) at a current sequencing depth. We also show that the relationship between saturated fold-enrichment and the number of sequenced tags may be extrapolated to estimate the sequencing depth that would be required to reach saturation for a lower fold-enrichment ratio. It will be important to examine how well such extrapolations describe saturation properties of much larger data sets that are likely to become available in the near future.

The fold-enrichment ratio for a particular binding position may depend on diverse factors, such as binding affinity or efficiency of chromatin extraction. As its relationship to the functional importance of binding positions is uncertain, the desired fold-enrichment ratio target should vary between experiments. When some functional binding positions are already known for a particular protein, the target enrichment ratio can be chosen based on examination of these positions in the initial sequencing data or with quantitative PCR. If a target enrichment ratio cannot be estimated from other sources, it can be specified relative to the maximum or median enrichment observed in the data set (e.g., to detect binding positions with enrichment fivefold below the maximum observed enrichment).

As more ChIP-seq data sets are generated, it will be important to analyze factors not considered here. These include sequencing biases associated with individual sequencing platforms and the stability of ChIP and input tag distributions between replicate experiments. Such data will likely improve binding prediction methods and allow better interpretation of the functional relevance of observed variability in fold-enrichment ratios of different binding positions. Finally, it will be important to assess whether the described techniques can be adapted for analysis of histone modifications or other widely distributed chromatin marks that do not fit the models of point binding patterns examined in this work.

METHODS

Data sets. The analysis of the NRSF binding was performed using tag data from Johnson et al.7. Raw tag information necessary for the analysis was only available for experiment no. 2. CTCF data was taken from Barski et al.13. The STAT1 binding was analyzed using the INF-γ–stimulated data set from Robertson et al.11.

Cross-correlation profiles. For each chromosome c, the tag count vector n′(x) was calculated to give the number of tags whose 5′ ends map to the position x on the strand s. Strand cross-correlation for a strand shift δ was then calculated as X(δ) = ∑n′(x)P(n′(x + δ), n′(x − δ)), where P(a,b) is the Pearson linear correlation coefficient between vectors a and b. C is the set of all chromosomes, Nc is the number of tags mapped to a chromosome c and N is the total number of tags.

Tag alignment and selection of informative tags. Sequence tags were aligned to human genome assembly (NCBI build 36, hg18) using BLAT16, with a minimum score threshold of 16, maximum gap of 4 and step size of 3. Tags aligning to multiple locations in the genome were discarded in this analysis.

Anomalous tag positions were identified as those with the number of mapped sequence tags (5′ ends) above the significance threshold defined by a Z-score of 10. All of the tags mapping to such an anomalous position (on either strand) were omitted before further analysis.

The resulting tag alignments were classified based on (i) the length of the alignment and (ii) the number of nucleotide differences (number of mismatches + total gap length). A given class of tags was accepted if adding these tags to the reference set significantly (Z-score > 6) increased the cross-correlation profile within the region ±20 bp around the cross-correlation peak of the reference set. The set of perfectly aligned tags (maximum length, no mismatches) was used as a reference set.

Detection of binding positions. For the WTD method, a binding score was calculated for all positions i in the genome as Swtd(i) = 2√(S0(i)−(p0 + n2)), where p0 and n2 are the number of 5′-end tag positions mapping to a positive strand within a distance of w upstream and downstream of a position i, respectively. Similarly, p2 and n2 correspond to the number of upstream and downstream tags mapping to the negative strand. Window size w = 200 bp was used for CTCF and NRSF, and w = 400 bp was used for STAT1. The window sizes were chosen to encompass the size of the average binding tag pattern (Supplementary Fig. 12). We found that this size can be estimated from the cross-correlation profiles (Fig. 1d and Supplementary Fig. 1) as the width of the cross-correlation peak at one-third of the peak height. Positions on the chromosome corresponding to nonunique tag alignment and mirror positions with respect to point i were excluded from score calculation. Binding peaks were determined as local maxima of Swtd(i).
MSP tag density profiles along each chromosome strand were calculated using Gaussian smoothing kernel with bandwidth corresponding to the 0.45 * σ\text{cc}, where σ\text{cc} is the width of the strand cross-correlation peak (Fig. 1d) at half width. The kernel bandwidth was selected for optimal coverage and accuracy of the method (Supplementary Fig. 19 online). A binding position was accepted when local maxima (peaks) of positive- and negative-strand density were found the distance of μ ± 20 bp, where μ is the size of the protected region for that protein (estimated from cross-correlation analysis). The peaks were required to be comparable in magnitude (based on a likelihood ratio test with a Z-score cutoff of 8).

For MTC, similar to WTD, the binding score was calculated as
\[ S_{\text{mc}}(i) = \rho / S_{\text{nc}}(i) + S_{\text{wi}}(i), \]
where \( \rho \) is the Pearson linear correlation coefficient between tag vectors \( v^+ \) and \( v^- \), such that \( v^+(k) \) is the number of 5'-end tag positions mapping to the positive strand in position \( i + k \), and \( v^-(k) \) is the number of 5'-end tag positions mapping to the negative strand at \( i - k \). The correlation is evaluated for \( k \in \{-w, \ldots, -w_0\} \cup \{w_0, \ldots, w\} \), where \( w_0 = 15 \) bp, and the values of \( w \) are the same as in the WTD method.

When using the methods described above, peaks within distance \( w \) of a larger peak were omitted (\( w = 200 \) bp for CTCF and NRSF, 400 bp for STAT1). The CSP method implementation provided by Johnson et al.\(^2\) was used, and the XSET method was implemented as described in Robertson et al.\(^13\).

Background tag density corrections. To normalize for background tag density in the analysis of NRSF binding, we adjusted the window tag counts described in the WTD and MTC methods by subtracting the weighted number of background tags occurring within that window. To account for differences between ChIP and background data set sizes, we multiplied the background tags by \( N_{\text{ChIP}} / N_{\text{Bg}} \), where \( N_{\text{ChIP}} \) and \( N_{\text{Bg}} \) are nonspecific sizes of ChIP and background data sets. The nonspecific size of the data set was determined as a number of data set tags outside of highly enriched regions: regions of 1 Kbp with the number of tags exceeding uniform (Poisson) density with \( P\text{-value} < 10^{-5} \). This type of weighting allows for proper estimation of the background density ratios when a large fraction of ChIP data set tags is concentrated within localized bound regions (this fraction for NRSF is 23%).

To reduce the impact of false-positives from large regions of systematically high background, subsequent calculation excluded regions of size \( 10^6 \) bp or larger where input tag counts are significantly larger (\( Z \)-score \( \geq 5 \)) than ChIP counts.

Statistical significance of detected positions. For a predicted binding position with score \( s \), the FDR was estimated as \( \frac{N_{\text{ChIP}}(s)}{N_{\text{ChIP}}(s) + \sigma_{\text{bg}}(s)} \), where \( N_{\text{ChIP}}(s) \) is the number of binding positions with score \( s \) or higher found in the real data set, and \( N_{\text{ChIP}}(s) \) is the number of binding positions found in a control data set. The FDR estimates of positions with scores above maximal score found in the control data set (that is, \( N_{\text{ChIP}}(s) = 0 \)) were assigned minimal FDRs found in the set of detected positions. Two types of control models were used: randomized models and a model based on the background (input) tag data.

Under a completely randomized model, control data were generated by randomly reassigning positions of the real (ChIP) data set tags. More restrictive randomization models maintained together tags that occurred within a distance \( d \) in the original data. Supplementary Table 3 shows results-based values of \( d \) ranging from 1 to 7. Ten randomized permutations of the complete data set were used for FDR calculations.

Under a background-based model, the control predictions were generated in the same way as predictions on real data, interchanging background (input) and ChIP data.

Sequence motifs position accuracy. Motif occurrence positions within the human genome were calculated using MAST\(^17\). High-scoring NRSF motif occurrences were determined using the position-specific matrix (PSSM) from Mortazavi et al.\(^14\). Positions with \( P < 4 \times 10^{-5} \) were chosen to match the number of motifs obtained in Johnson et al.\(^2\). For STAT1, GAS motif occurrences were determined using the PSSM from the TRANSFAC database\(^18\), with maximum \( P \)-value of \( 10^{-5} \). High-confidence CTCF motif positions were determined using the PSSM from Kim et al.\(^15\), with a \( P \)-value threshold of \( 4 \times 10^{-9} \). Spatial accuracy comparison using relaxed \( P \)-value thresholds is shown in Supplementary Figure 20 online.

The accuracy of the predicted protein-binding positions was analyzed based on the distances between identified positions and centers of high-scoring motif hits. Only binding positions occurring within 300 bp of a sequence motif instance were included in the analysis. The sign of the distance was adjusted according to the strand on which the motif hit occurred. Because the center of the motif hit may not represent a true center of binding (e.g., protected region), the distances to the motif were centered by subtracting the median distances. The centered distances were used in Figure 5b-d and Supplementary Fig. 11.

Sequencing depth analysis. To evaluate the stability of the identified set of binding positions on the set of tags, binding positions were predicted on randomly sampled subsets of the original tag data. Sampling was performed without replacement. The WTD method with an FDR of 0.01 was used to generate the predictions. A chain of subsampled data sets was generated by 15 successive random reductions of a data set by 10\(^2\) tags. A total of 100 such random chains were generated. The convergence of the MSER and depth predictions with the increasing number of chains is shown in Supplementary Figure 15.

We will use fractional agreement \( F(s, k) \) to refer to an average fraction of binding positions determined using a randomly sampled fraction of tags of size \( s \), that is also present (within 50 bp) in a set of binding positions determined using a tag subsample of size \( s \). The basic saturation curves (Fig. 6a) show the values \( F(s, x) \), where \( x \) is the number of tags sampled (\( x \)-axis), and \( s \) is the total size of the original data set.

To estimate the minimal fold-enrichment ratio of the identified binding positions over the background, we calculated the number of ChIP (\( n_F \)) and input (\( n_I \)) tags within 100 bp upstream or downstream of the identified position. The counts were used to estimate the 95% confidence interval of the fold-enrichment ratio based on a Poisson model with noninformative Bayesian prior\(^19\). As the background tag density is lower than the ChIP tag density, we also tested using larger window sizes in counting background tags (see Supplementary Fig. 17). Although such an approach should provide tighter enrichment confidence intervals, it appears to result in overestimation of enrichment folds relative to qPCR data.

We will use \( F(s, k) \) to refer to the fractional agreement after filtering both predictions to include only those binding positions with a lower bound of enrichment ratios \( > e \). The MSER for a data set \( x \) was calculated as the minimal value of \( e \) such that \( F(s, x, e) < 10^{-5} \). The relationship between \( \log_{10}(\text{MSER} - 1) \) and the size of the data set \( (x) \) was approximated using a linear model based on a least-squares fit.

Software availability. An implementation of the described methods is available as an R package and can be downloaded at http://compbio.med.harvard.edu/Supplements/ChIP-seq. The website will be updated periodically with new versions.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We would like to thank Dustin Schones and Keji Zhao for providing raw data and detailed descriptions for the CTCF data set, and Ali Mortazavi and Barbara Wold for providing sequence tag data for NRSF binding. This work was supported by grants from the National Institutes of Health to P.J.P. (U01HG004258, R01GM082798, U11RR024929).

Published online at http://www.nature.com/naturebiotechnology/ Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/
7. Bentley, D.R. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16**, 545–552 (2006).
8. Johnson, W.E. et al. Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl. Acad. Sci. USA* **103**, 12457–12462 (2006).
9. Qi, Y. et al. High-resolution computational models of genome binding events. *Nat. Biotechnol.* **24**, 963–970 (2006).
10. Peng, S., Alekseyenko, A.A., Larschan, E., Kuroda, M.I. & Park, P.J. Normalization and experimental design for ChIP-chip data. *BMC Bioinformatics* **8**, 219 (2007).
11. Robertson, G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
12. Smith, A.D., Xuan, Z. & Zhang, M.Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* **9**, 128 (2008).
13. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
14. Mortazavi, A., Leeper Thompson, E.C., Garcia, S.T., Myers, R.M. & Wold, B. Comparative genomics modeling of the NRSF/REST repressor network: from single conserved sites to genome-wide repertoire. *Genome Res.* **16**, 1208–1221 (2006).
15. Kim, T.H. et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231–1245 (2007).
16. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
17. Bailey, T.L., Williams, N., Misleh, C. & Li, W.W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369–W373 (2006).
18. Matys, V. et al. TRANSFAC: transcripational regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
19. Price, R.M. & Bonett, D.G. Estimating the ratio of two Poisson rates. *Comput. Stat. Data Anal.* **34**, 345–356 (2000).