nature biotechnology

PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls

Joel Rozowsky[1], Ghia Euskirchen[2], Raymond K Auerbach[3], Zhengdong D Zhang[1], Theodore Gibson[1], Robert Bjornson[4], Nicholas Carriero[4], Michael Snyder[1,2] & Mark B Gerstein[1,3,4]

Chromatin immunoprecipitation (ChIP) followed by tag sequencing (ChIP-seq) using high-throughput next-generation instrumentation is fast, replacing chromatin immunoprecipitation followed by genome tiling array analysis (ChIP-chip) as the preferred approach for mapping of sites of transcription-factor binding and chromatin modification. Using two deeply sequenced data sets for human RNA polymerase II and STAT1, each with matching input-DNA controls, we describe a general scoring approach to address unique challenges in ChIP-seq data analysis. Our approach is based on the observation that sites of potential binding are strongly correlated with signal peaks in the control, likely revealing features of open chromatin. We develop a two-pass strategy called PeakSeq to compensate for this. A two-pass strategy compensates for signal caused by open chromatin, as revealed by inclusion of the controls. The first pass identifies putative binding sites and compensates for genomic variation in the 'mappability' of sequences. The second pass filters out sites not significantly enriched compared to the normalized control, computing precise enrichments and significances. Our scoring procedure enables us to optimize experimental design by estimating the depth of sequencing required for a desired level of coverage and demonstrating that more than two replicates provides only a marginal gain in information.

With the advent of new high-throughput sequencing technologies (Helicos HeliScope, Illumina Genome Analyzer, ABI SOLiD, Roche 454), most genome-scale assays that previously could only be done cost-effectively using genomic tiling microarrays can now be performed using DNA sequencing. One of the most common uses of tiling microarrays is for performing ChIP-chip[1–3], a procedure involving immunoprecipitation of DNA associated with a protein of interest, labeling the resulting DNA, and then hybridizing it to a genomic tiling microarray. Early adaptations of ChIP sequencing (e.g., STAGE[4], ChIP-PET[5,6]) used Sanger-based sequencing, which generally provided limited tags and was expensive. ChIP-seq[7,8] involves sequencing millions of short tags from the immunoprecipitated DNA fragments. Although >100 ChIP-chip experiments were carried out during the pilot phase of the ENCODE project[9], almost all ChIP experiments in the scale-up to the whole human genome employ ChIP-seq. ChIP-seq is also being used extensively for the modENCODE project.

Short-tag sequencing platforms yield sequence reads of sufficient length to uniquely map most tags and their associated DNA fragments to the genome of interest. The Illumina Genome Analyzer platform, developed by Solexa, was the first truly high-throughput sequencing technology used widely for ChIP-seq. Each lane of data typically generates several million ~30-nt sequence tags. Mapping these tags against the genome, we can identify regions that are overrepresented in the number of mapped tags or fragments, which might correspond to genomic locations of transcription factor binding. However, there are a number of issues that make scoring more complicated. We have developed a general methodology for analyzing ChIP-seq data using two data sets—involving human RNA polymerase II (Pol II) and STAT1—sequenced more deeply than most published ChIP-seq data sets[7,8]. Pol II, a component of the general transcriptional machinery, and STAT1, a representative sequence-specific transcription factor, both bind primarily to punctate regions of DNA in what is typically called point-source binding. As an aid in determining experimental design, we further analyzed target identification as a function of sequencing depth (that is, saturation) and the number of biological replicates (independent biological samples) required.

RESULTS

Characteristics of ChIP-seq data

ChIP-seq data sets were generated for Pol II in unstimulated HeLa S3 cells (an immortalized cervical cancer–derived cell line) and for STAT1 in interferon-γ–stimulated HeLa S3 cells (STAT1 is induced when a cell is stimulated by interferon-γ). Matching sequenced input DNA control data sets were obtained for both stimulated and unstimulated cells. Although we chose to use input DNA as the control, we could have used a ChIP-seq with a different antibody (e.g., IgG) or a ChIP-seq sample under a different cellular condition (e.g., unstimulated STAT1 ChIP).

Signal maps for both HeLa S3 Pol II and STAT1 for a region on chromosome 22 are shown in the first and third tracks of **Figure 1a**.

[1]Molecular Biophysics & Biochemistry Dept., Yale University, PO Box 208114, New Haven, Connecticut 06520-8114, USA. [2]Molecular, Cellular & Developmental Biology Dept., Yale University, New Haven, Connecticut 06520, USA. [3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA. [4]Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA. Correspondence should be addressed to J.R. (joel.rozowsky@yale.edu) or M.B.G. (mark.gerstein@yale.edu).

Received 13 August 2008; accepted 3 December 2008; published online 4 January 2009; doi:10.1038/nbt.1518

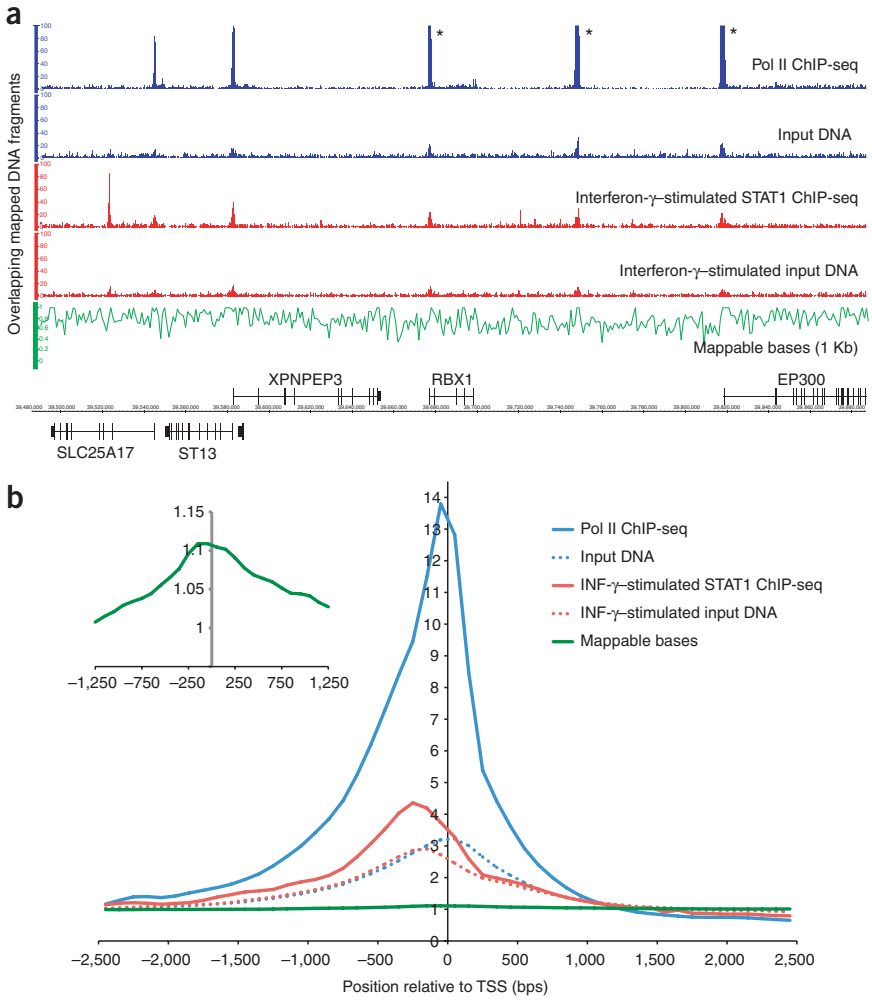


Figure 1 ChIP-seq signal profile maps. (**a**) The first and third signal tracks are plots of mapped fragment density for Pol II (in blue) and STAT1 (in red), respectively. The second and fourth tracks correspond to the input-DNA tracks for unstimulated (in blue) and interferon-γ–stimulated HeLa S3 cells (in red). The vertical axis for the first four tracks is the count of the number of overlapping DNA fragments at each nucleotide position (peaks in the top track indicated with a star have been truncated). The fifth track shows the fraction of uniquely mappable bases plotted in 1 Kb bins (in green). Many of the peaks in the Pol II and STAT1 tracks match corresponding peaks in the input-DNA controls, only some of which are enriched in their height relative to the control. (**b**) The signal for Pol II (solid blue line), STAT1 (solid red line) ChIP-seq and corresponding unstimulated (dashed blue line) and interferon-γ–stimulated (dashed red line) input-DNA controls are aggregated over regions proximal to all human CCDS transcription start sites (±2.5 Kb) plotted in 100-bp bins. There is significant enrichment for both transcription factors as well as the input-DNA controls over TSSs. The aggregated signal for the fraction of mappable bases is also plotted (green line) and there is a smaller but significant enhancement over TSSs (see insert where the vertical scale is from 0.95 to 1.15), though not as pronounced as the sequencing results.

The vertical axis is the count of overlapping mapped DNA fragments at each nucleotide position. Peaks (large numbers of overlapping mapped fragments) in this track correspond to regions of DNA where either Pol II or STAT1 has potentially bound in the HeLa S3 cell line being studied. Ideally the background to this experimentally generated signal map would be a randomly generated map with the same number of mapped fragments (that is, a uniform background distribution). If this were the case, peaks in the random background would follow Poisson statistics and could be computed either theoretically or by simulation. A peak threshold could then be set based on a false-discovery rate determined by the number of peaks from the background distribution compared to the actual data[7].

Unfortunately, the background distribution for a ChIP-seq experiment is not this simple[10]. There are multiple effects that contribute to the signal map from a ChIP-seq experiment. First, because sequence tags from certain genomic locations are not unique to the genome, sequenced reads from these regions would not be included, as they do not align uniquely to the genome. Thus the distribution of uniquely mappable bases in the genome is not uniform (**Fig. 1a**, fifth track).

Second, genomic DNA isolated from cells is in the form of chromatin. The structure of chromatin might bias the amount of DNA that is experimentally observable from different regions of the genome. There are also peaks in the signal maps for unstimulated and interferon-γ–stimulated HeLa S3 input DNA in the vicinity of promoters of known genes, which may correspond to regions of open chromatin[11] (**Fig. 1a**, second and fourth tracks). This can also be seen in **Figure 1b**, in which the signal maps have been aggregated proximal to transcription start sites (TSSs), that is, within ± 2.5 Kb, for all annotated Consensus Coding Sequences genes (CCDS gene annotations[12], uniformly agreed upon by Ensembl, NCBI and UCSC) in the human genome.

Thus the signal map of aligned fragments for a given transcription factor is actually the 'convolution' of a number of effects: the density of mappable bases in a region, the underlying chromatin structure and the actual signal from transcription factor binding. Therefore, some fraction of the peaks in the ChIP-seq signal map for a transcription factor might be due to the nature of the chromatin structure in those regions, that is, regions of open chromatin. To ascertain that the signal for any region is enriched owing to the presence of transcription factor binding, one must compare the signal against one from a control, such as a matching sequenced input-DNA experiment.

Mappability map of a genome sequence

A notable advantage of using tag-based sequencing instead of tiling microarrays for unbiased genomic experiments is that it is possible to cover a greater fraction of the genome. This is especially true for the more complex mammalian genomes that are comprised of almost equal amounts of repetitive and nonrepetitive sequence. In **Table 1** we compute the fraction of the genomes of four well-studied organisms (worm, fruit fly, mouse and human) that are assayable using either tiling arrays or tag sequencing–based technologies. For human we find that even though only 47.5% of the genome is nonrepetitive, 79.6% of the genome is uniquely mappable using 30 nucleotide (nt) sequence tags. Even for more compact genomes such as the worm's, which has much less repetitive sequence than the human genome, a substantial

Table 1 Genome mappability fraction

Organism	Genome size (Mb)	Nonrepetitive sequence		Mappable sequence	
		Size (Mb)	Percentage	Size (Mb)	Percentage
Caenorhabditis elegans	100.28	87.01	86.8%	93.26	93.0%
Drosophila melanogaster	168.74	117.45	69.6%	121.40	71.9%
Mus musculus	2,654.91	1,438.61	54.2%	2,150.57	81.0%
Homo sapiens	3,080.44	1,462.69	47.5%	2,451.96	79.6%

For four common model organisms—worm, fruit fly, mouse and human—we have determined the fraction of each genome sequence that is nonrepetitive as well as the fraction that is mappable using 30-nt sequence tags. The genome coverage achievable from genomic tiling arrays corresponds to the nonrepetitive fraction of a genome whereas the mappable coverage is what is achievable by tag-based sequencing approaches. We also determined that as the length of the sequence tags is increased beyond 30, the number of nucleotides in the genomes that are uniquely mappable is 2,452 Mb (79.6%) for 30-nt reads, 2,586 Mb (84.0%) for 40 nt, 2,669 Mb (86.7%) for 50 nt, 2,720 Mb (88.3%) for 60 nt and 2,750 Mb (89.3%) for 70 nt.

gain in coverage is achieved by using a tag sequencing–based approach (86.8–93.0% coverage with 30-nt tags). As next-generation sequencing technologies improve, longer sequence reads become possible. We find that the fraction of the human genome that is uniquely mappable increases from 79.6% to 89.3% as the length of the sequenced tag is increased from 30 to 70 nt.

We have developed code for efficiently indexing an entire genome and then determining at each nucleotide position the number of locations at which a sequence tag of length k appears in the entire genome (**Supplementary Notes** online). Analysis of the idealized achievable genome coverage from short-tag sequencing has previously been investigated[13]. The output from the code is a binary file for each chromosome, containing a table that maps from a position on the chromosome to the number of occurrences in the whole genome of the k-mer starting at that position. The mappability map that we construct only accounts for sequence tags that are located multiple places in the genome and that are identical; no mismatches are allowed. We have also investigated the effect of allowing for mismatches in the mappability map (**Supplementary Notes**).

We have determined the fraction of uniquely mappable nucleotide positions using genomic windows of 1 Kb (**Fig. 1a**, fifth track). We also have generated a profile of the fraction of alignable bases aggregated across all CCDS TSSs in the human genome and note a small enrichment in the fraction of alignable bases proximal to known TSSs (inset in **Fig. 1b**). The enhancement in mappable bases proximal to TSSs of genes is likely due to the increased complexity of DNA sequences in promoter regions.

Input DNA

To determine that a 'peak' in the signal map of DNA fragments actually corresponds to a site of transcription factor binding, it is necessary to show that the signal obtained is enriched compared to a matched control sample, such as input DNA isolated from the same cell line, under the same cellular conditions under which the ChIP experiment is performed, that is, HeLa S3 cells stimulated by interferon-γ for the case of STAT1. For input DNA, the distribution is not the 'flat' distribution one would expect from a random Poisson process (**Fig. 1a**, second and fourth tracks). There are more 'peaks' than would be expected from a random distribution. By analyzing the number of sites as a function of peak height, researchers have shown that this distribution cannot arise from a uniform background distribution[10]. There is a correlation between the locations of peaks present in the input-DNA signal map and the matching ChIP-seq results (**Supplementary Notes**).

Using the signal maps of DNA fragments, we created profiles of the aggregated signal maps proximal to the TSSs of well-annotated CCDS genes (**Fig. 2**). In addition we created profiles for both Pol II and STAT1. Although the aggregated profiles for HeLa S3 input DNA are not as pronounced as the aggregated signals from Pol II and STAT1, the input DNA under both conditions exhibits distinctive enrichments of signal proximal to TSSs. This again demonstrates that the peaks in the input-DNA signal do not arise from a random background distribution. We also note that the aggregated signals yield a higher definition profile with finer resolution than the aggregated ENCODE ChIP-chip profiles[9].

The aggregated profiles for the input-DNA samples proximal to TSSs are substantially more enriched than the relatively minor enrichment coming from the profiles of mappable bases (insert in **Fig. 1b**). This shows that although the mappability map is a component of the input-DNA signal, it only explains a relatively small portion of the enrichment proximal to TSSs. However, if one views the genome at a more coarse-grained level (averaged over 10 Kb windows), then we observe that when we scale the coarser-grained input-DNA signals by the fraction of mappable bases, the signal is more uniform than the signal before scaling. This shows that the fraction of mappable bases plays a substantial role in modulating the signal we observe.

PeakSeq: scoring ChIP-seq data

On the basis of these observations and our experience with scoring of ChIP-chip experiments[14], we have developed an approach, called PeakSeq, for scoring the results of ChIP-seq experiments by compensating for the mappability map and comparing these results against a normalized matching control data set. For computational efficiency we adopt a two-pass approach for scoring ChIP-seq data relative to a control data set. A schematic of the procedure is presented in **Figure 2**.

To accommodate the large data sets that are typically generated, we process all ChIP-seq data on a chromosome-by-chromosome basis. Using only uniquely mapping reads, we generate signal maps along each chromosome for the ChIP-seq data set as well as the matching control data set. Signal maps are generated by extending each mapped tag in the 3′ direction (as sequences are read from the 5′ end), to the average length of the DNA fragments in the sequenced DNA library ($\sim$200 bp). The signal map is then the integer count of the number of overlapping DNA fragments at each nucleotide position (**Fig. 2** (1)).

Motivated by the scoring procedure developed by others[7], in the first pass of our approach we focus on the ChIP-seq data set and identify regions or peaks in the ChIP-seq fragment density map that are substantially enriched compared to a simulated simple null background model. To capture some level of genomic variability (such as copy number variation[15,16]), we do this analysis on a segment-by-segment basis along each chromosome; segments are by default 1 Mb (**Fig. 2** (2)). Within each segment we use the mappability map to correct for the variation in mappability between segments. The candidate regions that are identified as potential DNA binding sites are not necessarily locations of transcription factor binding, as they may also be present in the input-DNA control. The first pass of the PeakSeq procedure acts as a pre-filter in which candidate regions are selected for comparison against the input-DNA control.

To compare the number of mapped tags to a potential binding site from the ChIP-seq sample compared to the control we need to normalize the control against the sample. We normalize the

background of the sample to the control by linear regression of the counts of tags from the control against the sample for windows (~10 Kb) along each chromosome. The slope of the linear regression α is used to scale tag counts from the control in the comparison with the ChIP-seq sample. Because windows that contain enriched peaks will increase the slope (conservatively overestimating the tag counts from the control), we introduce P_f—a parameter denoting the fraction of potential target regions that we exclude from the normalization procedure (windows that overlap excluded target regions are not used in the linear regression). We show the effect of the normalization procedure for two settings of this parameter ($P_f = 0$ and $P_f = 1$; **Fig. 2** (3) and **Supplementary Fig. 1**).

In the second pass of the procedure (**Fig. 2** (4), the ChIP-seq signals for putative binding sites are then compared against the normalized input-DNA control. Only regions that are enriched in the counts of the number of mapped sequence tags in the ChIP-seq sample relative to the input-DNA control are called binding sites. This comparison is analogous to the way enrichment is determined when validating ChIP 'hits' using quantitative (q)PCR. We compute the statistical significance using the binomial distribution. We also correct for multiple hypothesis testing by applying a Benjamini-Hochberg correction[17]. We report a ranked target list sorted by *Q*-value that also lists fold-enrichment values for each binding site. Comparison of potential target binding sites in the ChIP-seq sample against the input-DNA control accounts for the nonuniform background of a ChIP-seq experiment[10].

Application of PeakSeq to Pol II and STAT1 ChIP-seq data

We applied the PeakSeq procedure to the Pol II and STAT1 ChIP-seq data sets (we conservatively set $P_f = 0$ in the following analysis). We initially identified 73,562 and 123,321 potential binding sites for Pol II and STAT1, respectively. These represent the potential targets that are found to be enriched in the Pol II and STAT1 signal density maps compared to a simulated null random background. After comparing these target regions with the normalized input-DNA controls (unstimulated and interferon-γ–stimulated HeLa S3 cells), we found that only 24,739 and 36,998 of these regions are significantly enriched for Pol II and STAT1, respectively (using a false-discovery rate threshold

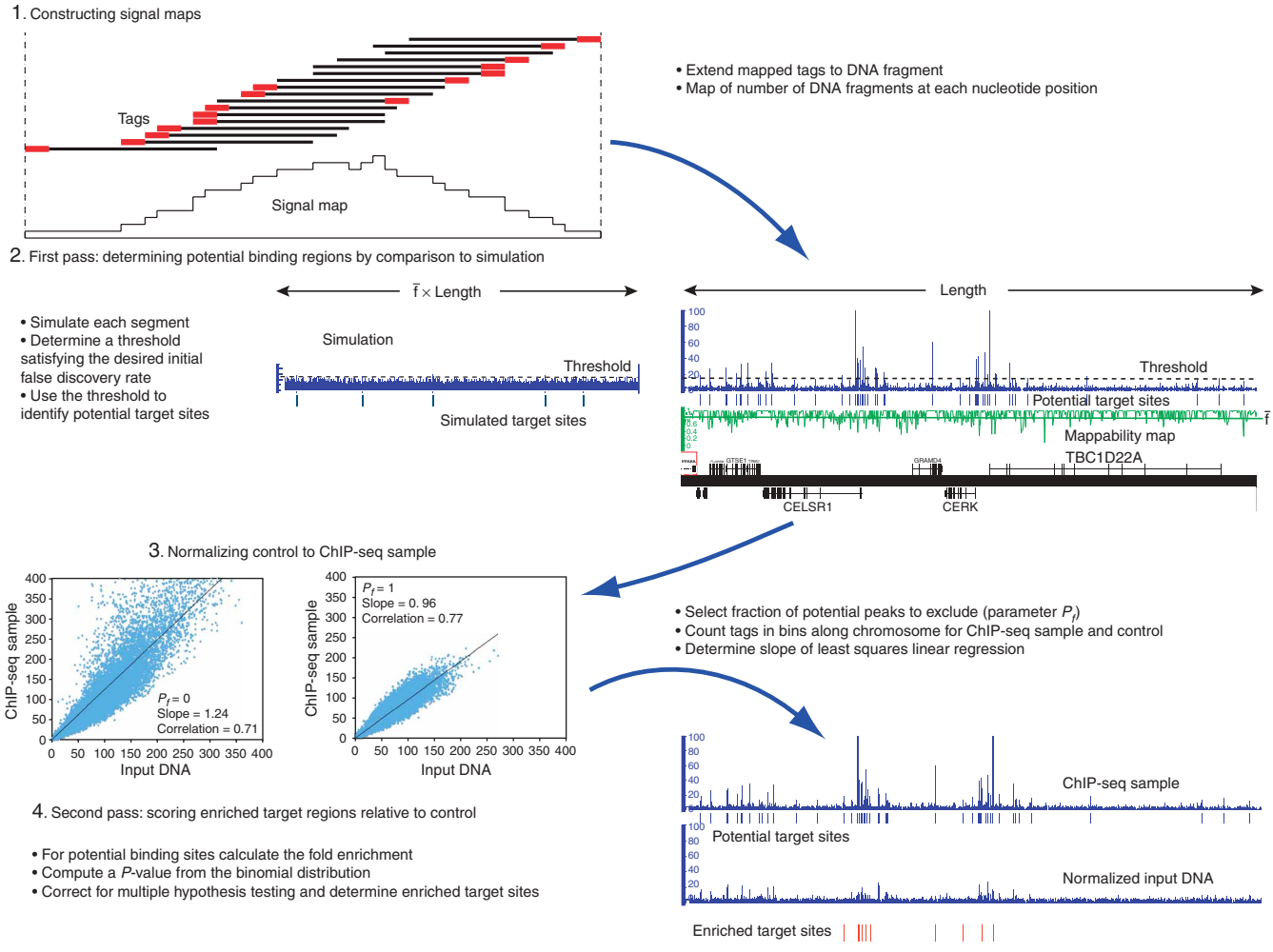


Figure 2 PeakSeq scoring procedure. (1) Mapped reads are extended to have the average DNA fragment length (reads on either strand are extended in the 3′ direction relative to that strand) and then accumulated to form a fragment density signal map. (2) Potential binding sites are determined in the first pass of the PeakSeq scoring procedure. The threshold is determined by comparison of putative peaks with a simulated segment with the same number of mapped reads. The length of the simulated segment is scaled by the fraction of uniquely mappable starting bases. (3) After selecting the fraction of potential target sites that should be excluded from the normalization, the scaling factor P_f is determined by linear regression of the ChIP-seq sample against the input-DNA control in 10-Kb bins. Bins that overlap the potential targets regions selected for exclusion are not used for regression. The fitted slopes as well as the Pearson correlations are displayed for P_f set to either 0 or 1. (4) Enrichment and significance are computed for putative binding regions.

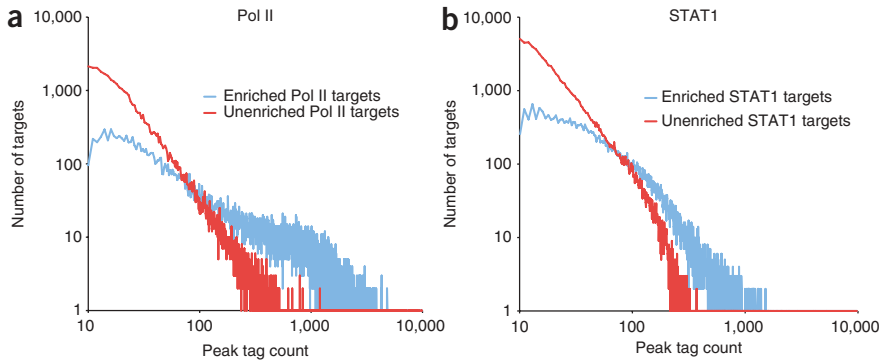


Figure 3 ChIP-seq target list scaling. (**a**,**b**) The distribution of target regions that are enriched (blue) relative to input DNA and those that are not (red) are shown on a log-log plot. The horizontal axis is the count of the sequence tags that are within a target peak, whereas the vertical axis indicates the number of target regions with that count. The left and right panels show the results for Pol II (**a**) and STAT1 (**b**), respectively.

of less than 0.05). We demonstrate how the number of target binding sites varies for a range of different false-discovery rate thresholds for each target list (**Supplementary Table 1** online).

We divided the putative targets identified for Pol II (**Fig. 3a**) and STAT1 (**Fig. 3b**) into targets that were enriched (blue) and those that were not significantly enriched (red). The horizontal axis, plotted on a log-log scale, is the count of the number of sequence tags overlapping a putative binding region. Both the enriched and unenriched (potential binding sites that are not enriched compared to the input-DNA control) distributions display an approximate power-law behavior. These plots are consistent with those generated for models of different background distributions[10]. However, the slope of the distribution of regions that are not enriched is steeper than that of the enriched distribution.

Comparison of PeakSeq results with published ChIP-chip and ChIP-seq data

We also compared the results obtained for Pol II and STAT1 ChIP-seq against matching pilot-phase ENCODE ChIP-chip data sets[9]. We found that 1,499 Pol II binding sites (averaging 275 bp) and 1,164 STAT1 binding sites (averaging 128 bp) are present in the 1% of the genome studied. Correspondingly, for ChIP-chip 1,000 Pol II sites, averaging 1,300 bp, and 395 STAT1 sites, averaging 507 bp, were identified (ChIP-chip experiments were performed using Nimblegen ENCODE tiling microarrays). We found that 321 (32.1%) of the ENCODE Pol II ChIP-chip sites were common to the matching ChIP-seq target lists. For the case of STAT1, many of the ChIP-chip target sites were independently tested for validation by qPCR[6]. We found that 106 of the 128 (83%) of the validated targets are common to the STAT1 ChIP-seq target list. By comparison, only 26 of the 282 (9%) regions that were not validated by qPCR are present on the ChIP-seq target list. In both cases, ChIP-seq was able to detect more binding sites (substantially more for STAT1) with higher resolution (that is, more localized binding sites). A representative example of the cytokine receptor locus on chromosome 21 comparing ChIP-seq and ChIP-chip for Pol II and STAT1 is shown in **Figure 4**.

We found that 21,750 of the 36,998 STAT1 ChIP-seq target sites (58.7%) are in common with the ChIP-seq results[7] (**Supplementary Notes** for details). When we ran the ChIPSeqMini[8] software on the Pol II ChIP-seq data produced in this paper, 9,467 target binding sites were identified with a median size of 1,939 bp. By comparison,

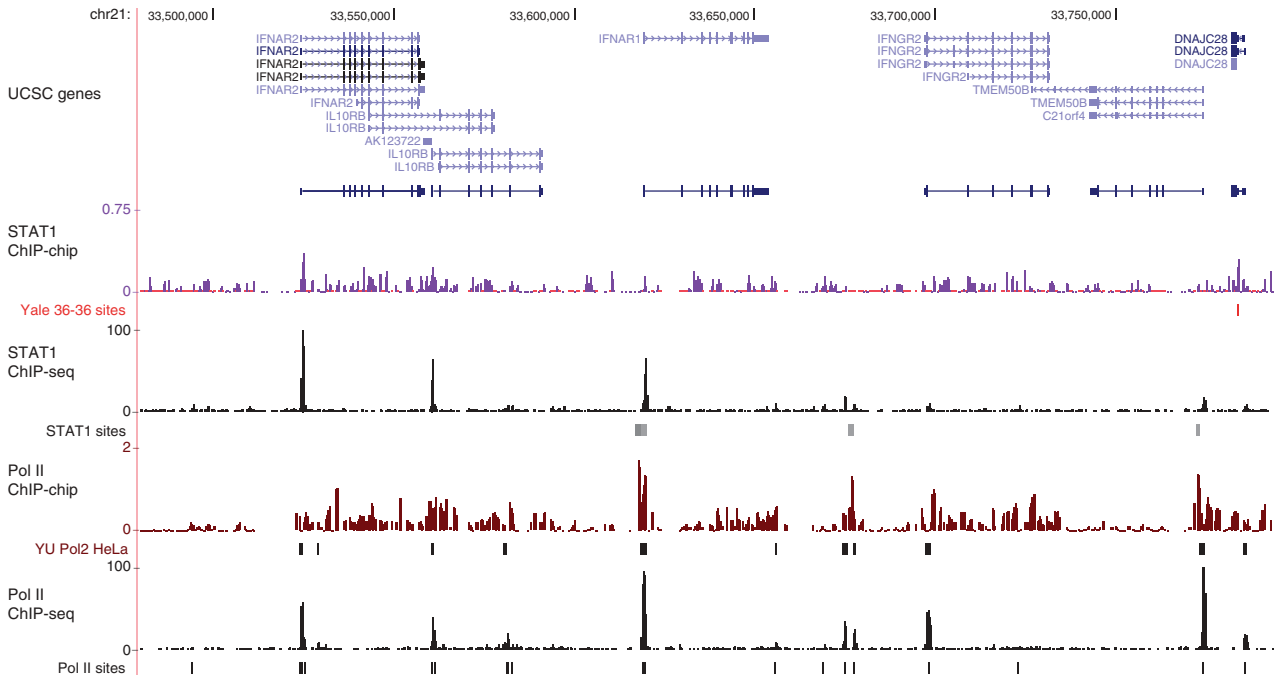
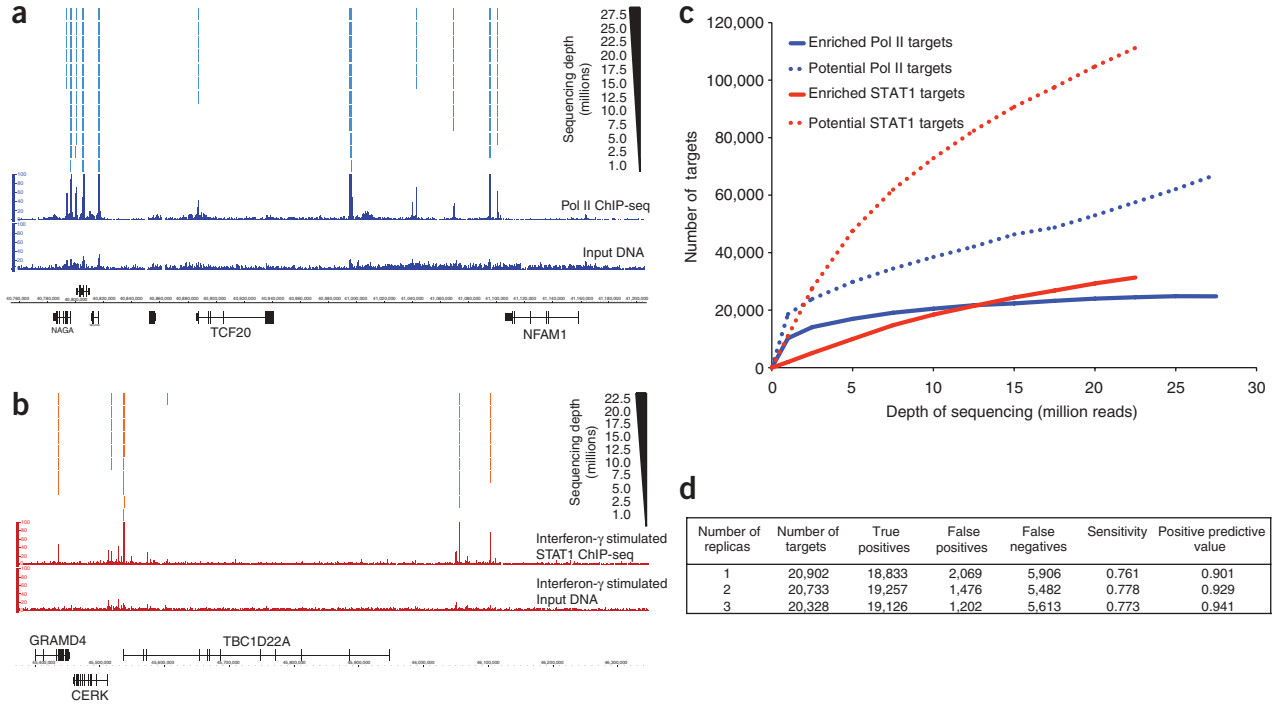


Figure 4 ChIP-seq versus ChIP-chip signal tracks and target binding sites for Pol II and STAT1. The ChIP-chip data were generated as part of the pilot-phase of the ENCODE project for 1% of the human genome. The region displayed is the cytokine receptor locus on chromosome 21. The ChIP-seq signal has a better signal-to-noise ratio and is higher resolution than the corresponding ChIP-chip data. Data were obtained from the UCSC Genome Browser[12].



d

Number of replicas	Number of targets	True positives	False positives	False negatives	Sensitivity	Positive predictive value
1	20,902	18,833	2,069	5,906	0.761	0.901
2	20,733	19,257	1,476	5,482	0.778	0.929
3	20,328	19,126	1,202	5,613	0.773	0.941

Figure 5 Depth of sequencing and value of replicates. (**a**) Fragment density signal tracks for Pol II and the input-DNA control as well as the target regions that are identified (significantly enriched) as a function of the number of mapped sequence reads. The same numbers of sequence reads are used for both sample and control. More prominent peaks are identified with fewer reads, whereas weaker peaks require greater depth. (**b**) Similar plot with STAT1 and matching interferon-γ–stimulated HeLa input-DNA control. (**c**) The number of putative Pol II (blue line) and STAT1 (red line) targets identified and the fraction for each of these that are enriched relative to input DNA as a function of the number of mapped sequence reads. Although the number of putative targets continues to increase for both Pol II and STAT1, the number of enriched targets begins to plateau. The number of Pol II targets appears to saturate faster than for STAT1 targets. (**d**) Summarized results of analyzing 9 million mapped Pol II ChIP-seq sequence reads using one, two or three biological replicates. We calculate sensitivity and positive predictive values using the targets identified with all the available sequence reads (~29 million uniquely mapped reads) as gold standard positives and the remainder as negatives. Only a marginal gain in positive predictive value at the cost of sensitivity is gained by using three biological replicates instead of two biological replicates.

PeakSeq identified 24,739 sites with a median size of 841 bp. Using the same data and the same hardware, PeakSeq ran in <8 min whereas ChIPSeqMini required 459 min to run. All the regions identified by ChIPSeqMini overlap regions called by PeakSeq, which is consistent with ChIPSeqMini using a more restrictive threshold in calling peaks, even though the regions identified are broader. The tallest peaks identified in the initial peak-calling pass of PeakSeq tend to all be significantly enriched compared to the input-DNA control.

Implications for the optimal design of ChIP-seq experiments

To investigate the depth of sequencing required to saturate the number of identifiable target binding sites, we carried out the following analysis. For both the Pol II and STAT1 ChIP-seq data sets, we shuffled the mapped sequence reads to remove any biases due to the variation between biological replicates or different flow-cell lanes. The sequences for each of the input-DNA controls were similarly shuffled.

Initially, for each transcription factor one million reads were selected from both the randomly ordered sample and control sequences. These data sets were scored, identifying both potential targets for binding as well as the subset of these that are enriched compared to the control. For increasing numbers of reads, we scored putative and enriched target sites (larger sets of reads are inclusive of smaller sets (**Fig. 5**)).

We plot the number of identified putative targets (dashed lines) and enriched targets (solid lines) for both Pol II (blue) and STAT1 (red) (**Fig. 5c**). The number of Pol II targets saturates as a function of sequencing depth. The number of targets for Pol II appears to approach an asymptotic value of ~25,000 target regions. Curiously, the number of targets identified for STAT1 initially climbs much slower than those for Pol II; however, the number of targets continues to rise and only starts to show signs of plateauing once 22.5 million reads have been analyzed. This is consistent with the larger proportion of Pol II targets that show higher levels of enrichment compared with STAT1. We note that for both Pol II and STAT1 the number of putative target regions continues to increase significantly as a function of sequencing depth. This analysis implies that the set of identified sites is approaching the total number of target sites (or that the total number of sites is saturating). However, formally this is not a proof of saturation as, in principle, there could be another regime after some critical number of sequence reads where it is possible to identify a novel class of enriched regions (that is, broad binding domains).

In **Figure 5a**,**b** are the target binding sites that are enriched as a function of the number of mapped sequence reads. The most prominent peaks are identified with only one million sequences: however, smaller peaks are only called when more sequences are included.

Biological replicates in high-throughput genomic experiments are carried out to achieve two disparate objectives: to ensure that experiments are reproducible and to quantify the biological variation between samples in an experiment. As part of the pilot phase of the ENCODE project[9], it was decided that three biological replicates were necessary for ChIP-chip experiments to ensure reproducibility.

What is the optimal number of biological replicates necessary for a ChIP-seq experiment?

To quantify the gain in the number of enriched target binding sites identified by adding additional biological replicates, we carried out the following analysis using the ~29 million uniquely mapped sequence reads for Pol II ChIP-seq that were generated using three independent biological preparations. Using the same total number of mapped reads, we separately analyzed the results using sequences coming from increasing numbers of replicates. Care needs to be taken to ensure appropriate randomization of reads and permutations of replicates from flow-cells or lanes to avoid biases.

Using the target list identified from all the available Pol II ChIP-seq data as a gold standard set of positives (and all other regions as negatives), we compared the target lists identified using 9 million reads from one, two or three biological replicates, and computed the sensitivity and positive predictive value for each target list. We then calculated the average sensitivity and positive predictive value for one replicate (averaged over all three), two replicates (averaged over all three pairs) and three replicates (results summarized in **Fig. 5d**). There was some gain in sensitivity and positive predictive value when the number of biological replicates was increased from one to two. However, there is no further increase in sensitivity when using three replicates and the positive predictive value increases only marginally. In a ChIP-seq experiment it is necessary to perform at least two replicates to ensure that experimental results are reproducible (to identify a failed experiment); however, it is not clear that there is a substantial information gain beyond two independent replicates. Consequently the ENCODE consortium currently requires >90% agreement between target lists from biological replicates.

DISCUSSION

We have demonstrated that there are two main observed effects that modulate the ChIP-seq signal profile: the genomic variation in the fraction of mappable sequence tags and differences in chromatin accessibility as evidenced in sequenced input DNA control experiments. These two effects can be contrasted with the two main biases that affect tiling array ChIP-chip experiments: probe-to-probe hybridization differences and cross-hybridization[18]. In reality we use the same input-DNA control as in ChIP-chip experiments and the same chromatin features should be present when observing the signal for the single-channel input-DNA control. This is typically not apparent as the other two effects, which are more pronounced, obscure this signal. The effect of the input DNA is the same for ChIP-chip and is normally scaled out when the ChIP-chip signal is scored in the typical two-channel fashion.

One might think that ChIP-seq experiments are immune to the probe-specific effects and cross-hybridization that both occur for ChIP-chip; these two effects can be contrasted against the genomic variation in mappable sequence tags, which plays a similar role in the analysis of ChIP-seq data sets. One fundamental difference between ChIP-chip and ChIP-seq is that the signal for ChIP-chip is a continuous-valued fluorescent intensity for each oligonucleotide probe, whereas the signal for ChIP-seq is a discrete integer-valued count of the number of mapped tags in a genomic region. This affects the analysis and the type of statistics used. Motivated by the way ChIP-chip experimental data are analyzed, we have developed an approach for scoring ChIP-seq data accounting for variation in mappability and input-DNA controls. Initial analyses of ChIP-seq[7,8] experiments did not account for these effects and target regions were not scored relative to an appropriate control experiment.

Although we have initially implemented our methodology for use with tag sequence data from the Illumina Genome Analyzer platform, it should be relatively straightforward to adapt it for use with other high-throughput sequencing platforms. We have developed the PeakSeq approach to identify peak regions in ChIP-seq data sets that correspond to sites of transcription factor binding. Although we have used input DNA as a control in this paper, other controls can be used (e.g., unstimulated STAT1 ChIP or IgG). We have also shown that separate input-DNA control samples are needed for different cellular conditions even for the same cell line (**Supplementary Notes**). Although this approach has been developed and calibrated to identify sites for more punctate point-source binding of transcription factors or proteins to DNA, it can also be used to identify broader regions of binding (such as histone modifications) that show significant enrichment relative to control. However, a more detailed procedure will be necessary for identifying extended regions of binding. A notable feature of our analysis methodology is that statistical quantities such as false-discovery rates and *P*-values are based on the number of target regions identified rather than the number of enriched nucleotides. In our approach, we treat sequence tags mapping to the forward and reverse strands equally (tags are sequenced from the 5′ ends of DNA fragments). One could compare the relative orientation of these reads in each target region compared to what would occur by chance.

For certain transcription factors, ChIP-seq surpasses ChIP-chip for identifying sites of transcription factor binding[7]. By analyzing the signal data between ChIP-seq and ChIP-chip data we show that ChIP-seq data give finer resolution and a greater signal-to-noise ratio (**Fig. 4**). ChIP-seq also achieves substantially greater coverage of the genome of interest as compared with ChIP-chip using tiling arrays, especially for the larger mammalian genomes. The fold enrichment determined for ChIP-seq typically shows a significantly greater range than does ChIP-chip; this is understandable as effects such as cross-hybridization tend to reduce the fold enrichment from a tiling array approach. Although ChIP-seq appears to significantly outperform ChIP-chip it is not clear that this will be the case for all transcription factors and chromatin modifications, especially those that bind to broader genomic regions where it might be necessary to sequence extremely deeply to achieve better results than from tiling arrays.

METHODS

Generation of Illumina tag sequencing data. For this paper we generated two deeply sequenced ChIP-seq data sets for antibodies against both human RNA polymerase II and STAT1 performed in the HeLa S3 cell line. For Pol II using the results from 24 lanes of Illumina sequencing data, we obtained >29 million mapped reads for the Pol II ChIP-seq as well as a matching 29 million reads for a control sample of HeLa S3 input DNA. STAT1 ChIP was performed in HeLa S3 cells that had been stimulated using interferon-γ, producing 26 million mapped reads for IFN-γ–stimulated STAT1 ChIP DNA. We obtained 24 million mapped reads for a matching control of IFN-γ–stimulated HeLa S3 input DNA. See **Supplementary Notes** for further details of Methods. A complete breakdown of the ChIP-seq reads generated is summarized in **Supplementary Table 2** online.

Illumina data analysis pipeline. Raw image files from the Illumina Genome Analyzer machine were transferred to the Yale biomedical high-performance computing cluster. Data are processed following the recommended Illumina pipeline. Raw images are first processed using the Firecrest software package. Base-called sequences with confidence metrics are obtained using the Bustard software. Gerald is the last part of the pipeline. It uses the program ELAND to align the short-sequence reads against the genome of interest allowing for up to two possible mismatches. By default ELAND only gives the locations for reads that align uniquely to the target sequence; however, with modified parameters ELAND can report the locations for reads that map to multiple locations. For the human ChIP-seq and control data sets, we aligned the sequence reads to the

latest build of the human genome (hg18/NCBIv36) obtained from the UCSC Genome Browser[12]. We excluded the random unassembled contigs.

Alignment of tag sequences. After a typical analysis pipeline, flow-cell images are analyzed, yielding base called sequences with confidence scores, which are then aligned against the appropriate genome. The standard Illumina pipeline uses the program Eland for aligning sequence tags, although a number of other applications have been developed for the same purpose, for example, Maq[19], Rmap (A. Smith, Univ. of Southern California and Z. Xuan, Cold Spring Harbor Laboratories, personal communication), SOAP[20]. When aligning sequence reads against the genome, reads aligning to multiple locations are typically excluded, as they are ambiguous (**Supplementary Table 2** to see the proportion of sequence reads generated that map to multiple locations in the human genome). However, such exclusion results in portions of the genome, including highly repetitive sequences or recent segmental duplications, that are not alignable and thus not assayable.

Although the algorithm we have developed by default only uses sequence tags that map uniquely, it would be a relatively straightforward modification to use tags that map to multiple locations by capping the number of locations to which a tag is allowed to map and by weighting tags by a factor dependent on the number of locations to which it maps or randomly selecting one of the locations. Many of the mapping algorithms including Eland and Maq have options allowing for the aligning of sequence reads to multiple locations in the genome. The effect of including sequence reads that map to multiple locations in the scoring is discussed here.

PeakSeq: first pass identification of potential target sites. Once fragment density maps have been created for both the sample and control data sets, we initially focus on the sample density map. Each chromosome is subdivided into segments of length $L_{segment}$ (typically 1 Mb) for analysis. Depending upon the genome being analyzed one might select a different size for these segments, for example, smaller segments for more compact genomes. For each segment the number of fragments that align to that segment are counted, N_{reads}. In addition, by using the mappability map for the genome of interest, we can precompute the fraction of uniquely alignable bases in that segment, f. Using these two parameters, N_{reads} and f, we can perform a computational simulation by randomly generating N_{reads} aligned DNA fragments in a scaled segment of length $f \times L_{segment}$ (**Fig. 2**, and below for details of the simulation). To perform an accurate simulation this is done multiple times and the results of these simulations are averaged.

Using a height threshold we can determine all the contiguous regions that are above this threshold in the sample fragment density map. Regions above threshold that are separated by genomic distances less than the average fragment length ($\sim$200 bp) are merged. This is similar to the maxgap/minrun approach that is commonly used for analyzing ChIP-chip tiling array data[14,21]. For the same threshold we can determine the number of merged regions above the threshold in the simulation. For each threshold the fraction of false positives is calculated from the ratio of the estimated number of false positives above threshold from the simulation divided by the number of regions above threshold for the ChIP-seq sample. By tuning the threshold used, we can set an initial first pass false-discovery rate (the false-discovery rate for the final list of target binding sites will be determined after comparison with the control sample). The threshold is set independently for each segment of each chromosome, which accommodates for genomic variability along each chromosome due to, for example, structural variation[15,16].

Using this thresholding procedure, we obtain candidate sets of peak regions (that is, putative binding sites) from each chromosome that are significantly enriched compared to a null random background for each segment. However, some of these regions might be due to underlying peak signal that is present in the control sample. Thus, to determine whether each of these putative peaks is actually bound by the transcription factor, we need to show that the number of mapped fragments from the sample data set is significantly enriched compared the control input-DNA data set.

Estimation of false positives by simulation in the first pass of PeakSeq. For each segment of length $L_{segment}$, a computational simulation of N_{reads} tag sequences is performed using the scaled segment length $f \times L_{segment}$ (the length is scaled by the fraction of uniquely mappable bases in the segment). N_{reads} tag sequences are randomly placed along the $f \times L_{segment}$ segment length. The same thresholding procedure is then followed for determining false positives from the simulated data (**Fig. 2**). The simulation is performed multiple times and the number of false positives is averaged over the different simulations.

We only use a simple background null distribution (Poisson) for each segment, rather than the more-complicated background model[10], during the first pass of the PeakSeq procedure. This is because we are trying to identify a candidate list of potential target regions using a relatively liberal threshold. The nonuniformity of the background will be accounted for in the second pass when counts of mapped fragments for putative binding sites are compared against the input-DNA control. The control captures the actual background distribution, which we do not need to model explicitly. If we were scoring the ChIP-seq data without a reference control then the nonuniformity of the background would have to be modeled explicitly.

Application of the mappability map to PeakSeq scoring. The mappability map is initially constructed for each base pair in the genome (**Supplementary Notes**), counting the number of locations to which a sub-sequence starting at that position, typically of length 30 nt, aligns. Using this we can generate a coarse-grained map of the fraction of uniquely mappable bases (corresponding to tags starting at those base pair locations) in a segment (that is, window) of a given size. In the paper we have used coarse-grained maps for segments of size 1 Kb for illustrative purposes in **Figure 1a**,**b**. In addition we have generated a coarse-grained mappability map using larger 1 Mb segments. As part of the first pass filtering in the PeakSeq scoring procedure (**Fig. 2** (2)), we determine a peak-height threshold determined for each 1-Mb segment in the genome. For a segment, the threshold is determined by comparison against a simulated null background using the same number of tag reads randomly mapped onto a region of length corresponding to the number of uniquely mappable bases in that 1-Mb segment (that is, the fraction of mappable bases multiplied by the segment length).

PeakSeq: normalization of control to ChIP-seq sample. Before this comparison can be made the control data set has to be appropriately normalized to the sample data set. Naively one could use matching data sets with the same number of mapped reads by removing reads from the larger data set. This is not the correct way to perform the normalization, as it is overly conservative. The sample data set is composed of a portion of mapped reads that come from the background distribution whereas the remainder arises from peak regions that are genuine binding sites. Mapped reads that are part of genuine binding sites would incorrectly skew the apparent parity achieved by simply using the same number of mapped reads between sample and control. We actually want to normalize the control data set against the background component of the sample data set. To reduce the effects of peaks in the normalization, we divide each chromosome into short segments (length $\sim$10 Kb) and perform the normalization using all segments that have at least one mapped fragment. We would like to exclude segments from the normalization procedure that contain peaks corresponding to binding sites; however, we do not want to exclude all putative binding sites identified in the first pass of the procedure as this would exclude segments that contain peaks that are present in both the sample and control background distributions. Thus we introduce a parameter, P_f, which is the fraction of the peaks (ranked by peak height) that should not be included in the normalization procedure. If $P_f = 0$ then no peaks are excluded and all segments are used for the normalization, whereas if $P_f = 1$, only segments that do not overlap any candidate peak are used for normalization. Using this procedure all the included segments contribute equally when computing the normalization factor rather than allowing the peaks to dominate.

For each segment (indexed by s) not overlapping the P_f fraction of putative peaks identified in the first pass, we count the number of mapped tags per segment for both the sample, N_s^{sample}, as well as the control, $N_s^{control}$. Chromosome by chromosome we perform least-squares linear regression between these two sets of counts, $N_s^{control}$ and N_s^{sample}. The slope of the regression is then a scaling factor, α, between the number of counts from the control and the sample of interest. In general, setting $P_f = 0$ will be more conservative because

both the slope α and the normalized counts of tags from the control for each potential target binding site will be larger and thus fewer regions will be deemed enriched relative to the control (**Fig. 2** (3)).

PeakSeq: second pass scoring target sites relative to control. For each of the putative binding regions, indexed by r, we can now count the number of mapped fragments that overlap the region from both the sample data set, N_r^{sample} as well as the number from the control data set, $N_r^{control}$. We appropriately normalize the count from the control by multiplying by the scaling factor computed above. For each putative site we first compute the fold enrichment, that is, the ratio of the number of mapped reads from the sample N_r^{sample} over the scaled number of mapped reads from the control, $\alpha \times N_r^{control}$. The fold enrichment is the signal normally computed for a transcription factor binding site, which should be proportional to the occupancy number for the binding site (the fraction of cells in the experiment that have the transcription factor bound at this site). Using the binomial distribution we can calculate a P-value of the significance of the region's enrichment in the number of fragments from the sample as compared to the scaled number from the control (because the scaled number is not in general an integer, this number is rounded up to the nearest integer value). The null hypothesis is that there is no enrichment.

As is typical in high-throughput experiments that generate a large number of results, corrections need to be made to account for multiple hypothesis testing. Due to the large number of statistical tests being performed, for any P-value threshold used some number of false positives (potentially many) will arise by random chance. Following a standard approach for the analysis of large-scale experiments, we employ a false-discovery[22–24] based approach using a Q-value[22–24]. A Bonferroni-type correction for multiple hypothesis testing is typically overly conservative so we choose to use a Benjamini-Hochberg correction[17].

Statistical tests. To determine whether a given putative target region r is enriched in the number of mapped tags from the ChIP-seq sample compared to the normalized input-DNA control, we calculate the P-value from the cumulative distribution function for the binomial distribution, which corresponds to summing the tail of the distribution. The cumulative distribution function is given by

$$F(k, n, P) = \sum_{j=0}^{\lceil k \rceil} \binom{n}{j} P^j (1-P)^{n-j}$$

where $k = \alpha \times N_r^{control}$ is the scaled number of sequence tags overlapping the target region from the input DNA, N_r^{sample} is the number of tags from the sample ($n = k + N_r^{sample}$) and $P = 0.5$, which is the probability under the null hypothesis that tags should occur with equal likelihood from the sample as from the control. Once nP is sufficiently large, the binomial distribution can be accurately approximated by a normal distribution

$$Norm(nP, nP(1-P))$$

with mean nP and variance $nP(1–P)$.

Correcting for multiple testing. We follow Benjamini and Hochberg[17] in adjusting our P-value to correct for multiple testing. All the target regions that are tested for significance are ranked by P-value from most significant to least significant. Then for each region the Q-value is then given by

$$Q - value = P - value \times \frac{Count}{Rank}$$

where *Count* is the total number of regions tested. Enriched target regions are then selected using a Q-value threshold rather than a P-value threshold.

PeakSeq software. The scoring procedure has been implemented in C and Perl and the source code is publicly available for download (http://www.gersteinlab.org/proj/PeakSeq/).

Scoring ChIP-seq data including reads that map to multiple locations. To investigate the effect of only including uniquely mapping reads, we selected a single lane of Pol II ChIP-seq and input-DNA data and included reads that aligned to at most ten distinct locations in the genome allowing for up to two possible mismatches per read. Alignments were performed using Eland. For reads that align to multiple locations, we randomly selected one of those locations. Scoring this data using the same PeakSeq procedure outlined below, we find that the number of binding sites identified increases by 17% compared to only using reads where the best match is unique. Thus we can use PeakSeq to score the reads that map to multiple locations; however, these binding sites are inherently ambiguous due to the nature of these sequences. Some of these regions will correspond to legitimate sites of factor binding to DNA. These results are available for download from http://www.gersteinlab.org/proj/PeakSeq/.

Analysis of ChIP-seq data from biological replicates. To appropriately compare sequence reads from biological replicates, we subdivided the data from each of the three different biological replicates (the sequences were also randomly permuted for each replica). We first selected 9 million reads from each of the three replicates (only 8.1 million reads were available for analysis from the third biological replica). Each of these data sets was scored against 9 million reads randomly selected from the input-DNA control (the same 9 million control reads were used as a control for each analysis). Second, we randomly selected 4.5 million reads from each of two different biological replicates, which were then combined to form 9 million reads and these were scored against the 9 million control reads as before. This was done for all three combinations of selecting two-of-three pairs of samples. Lastly, we selected 3 million reads from each of the three replicates, which were combined and scored against the sampled control data set. This type of analysis can be generalized for the comparison of more than three replicates.

Accession numbers. Raw and aligned sequence reads for all data sets have been deposited at GEO: GSE12781 (Pol II) and GSE12782 (STAT1).

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

ACKNOWLEDGMENTS

This work was done with support by grants from the National Institutes of Health (NIH) and made use of the Yale University Life Sciences Computing Center (NIH grant RR19895). We acknowledge Mike Wilson's assistance with submission of data to GEO.

AUTHOR CONTRIBUTIONS

J.R. conceived and developed the scoring methodology, analyzed the data presented in the paper and wrote the manuscript. G.E. generated the experimental data. R.K.A. assisted with the analysis in the paper as well as editing the manuscript. Z.D.Z. was involved in the conceptualization of the scoring methodology. T.G. assisted in the coding of the PeakSeq scoring procedure. R.B. and N.C. developed the code for generating indexed mappability maps of a genome and assisted with analysis. M.S. helped conceive of the scoring methodology and with the editing of the manuscript. M.B.G. also helped conceive of the scoring methodology as well as supervised the analysis and writing of the manuscript.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
2. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
3. Horak, C.E. & Snyder, M. ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol.* **350**, 469–483 (2002).
4. Kim, J. *et al.* Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat. Methods* **2**, 47–53 (2005).
5. Wei, C. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**, 207–219 (2006).
6. Euskirchen, G.M. *et al.* Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res.* **17**, 898–909 (2007).
7. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
8. Johnson, D.S. *et al.* Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).

9. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
10. Zhang, Z.D. *et al.* Modeling ChIP sequencing in silico with applications. *PLoS Comput. Biol.* **4**, e1000158 (2008).
11. Giresi, P.G. *et al.* FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17**, 877–885 (2007).
12. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
13. Whiteford, N. *et al.* An analysis of the feasibility of short read sequencing. *Nucleic Acids Res.* **33**, e171 (2005).
14. Zhang, Z.D. *et al.* Tilescope: online analysis pipeline for high-density tiling microarray data. *Genome Biol.* **8**, R81 (2007).
15. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
16. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
17. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
18. Royce, T.E., Rozowsky, J.S. & Gerstein, M.B. Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics* **23**, 988–997 (2007).
19. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
20. Li, R. *et al.* SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
21. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
22. Storey, J. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* **64**, 479–498 (2002).
23. Storey, J. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.* **31**, 2013–2035 (2003).
24. Gibbons, F.D. *et al.* Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization. *Genome Biol.* **6**, R96 (2005).