A comparative analysis of genome-wide chromatin immunoprecipitation data for mammalian transcription factors

Hongkai Ji, Steven A. Vokes¹ and Wing H. Wong²*

Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA, ¹Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA and ²Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305, USA

Received July 13, 2006; Revised September 26, 2006; Accepted September 28, 2006

ABSTRACT

Genome-wide location analysis (ChIP-chip, ChIP-PET) is a powerful technique to study mammalian transcriptional regulation. In order to obtain a basic understanding of the location data generated for mammalian transcription factors and potential issues in their analysis, we conducted a comparative study of eight independent ChIP experiments involving six different transcription factors in human and mouse. Our cross-study comparisons, to the best of our knowledge the first to analyze multiple datasets, revealed the importance of carefully chosen genomic controls in the de novo identification of key transcription factor binding motifs, raised issues about the interpretation of ubiquitously occurring sequence motifs, and demonstrated the clustering tendency of protein-binding regions for certain transcription factors.

INTRODUCTION

Genome-wide chromatin immunoprecipitation (also known as location analysis) is a powerful technique to identify and locate mammalian transcription factor binding regions at a resolution of 0.5–2 kb (1–4). Combined with downstream sequence analysis (5–9), this technology has the potential to provide a detailed characterization of structures [i.e. identities and locations of 6–30 bp long transcription factor binding sites (TFBS)] and functions of mammalian cis-regulatory elements. Applying this technology to mammalian genomes is still at an early stage; hence, systematic understandings about the data itself are limited, as is our knowledge about potential issues in their analysis and interpretation. Because of this, it is still not clear as to which types of computational analyses are generally useful, how they should be performed and how the results should be interpreted. To help clarify these issues, we performed a comparative analysis of eight independent ChIP experiments in human and mouse (Table 1). These experiments involve six different transcription factors [Gli, Estrogen Receptor (ER), p53, Oct4, Sox2 and Nanog] and three different technological platforms (ChIP-chip on Agilent tiling arrays, ChIP-chip on Affymetrix tiling arrays and ChIP-PET). Through the cross-study comparisons, we show that

(i) Current ChIP-chip and ChIP-PET technology are sufficient for the unambiguous de novo identification of transcription factor binding motifs from mammalian genomes. In all experiments with transcription factors containing biologically validated motifs, the signal-to-noise ratio provided by ChIP data was strong enough to support de novo motif discovery without the use of either cross-species information or the co-localization properties of 6–30 bp long TFBSs. This provides confidence in applying current ChIP technology to define novel mammalian transcription factor binding motifs.

(ii) Certain sequence patterns can be ubiquitously identified in transcription factor binding regions even after repeat-masking. Detection of these sequences therefore needs to be interpreted with caution.

(iii) Methods for generating matched genomic controls are critical for defining the transcription factor binding motifs that are of main interest for individual studies.

(iv) Rather than being randomly distributed, the binding regions of certain transcription factors have pronounced clustering tendencies. Unlike the clustering of 6–30 bp long TFBSs within a cis-regulatory module (CRM) 100 bp to 1 kb in length which had been studied extensively before (10–14), the tendency discussed here is the clustering of multiple 0.5–2 kb binding regions (potentially multiple CRMs) within a range of 1–100 kb. This tendency may be used to improve computational predictions of target genes for relevant transcription factors.

*To whom correspondence should be addressed. Tel: +1 650 725 2915; Fax: +1 650 725 8977; Email: whwong@stanford.edu

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
The study here represents the first cross-platform, multifactor analysis of genome-wide ChIP data. Our results revealed common characteristics of such data for mammalian transcription factors and provide guidelines for their future analysis.

**MATERIALS AND METHODS**

**Data preparation**

To conduct the comparative study, we collected ChIP-chip data for Gli (mouse, GEO accession no. GSE5683) (S.A. Vokes, H. Ji, S. McCuine, T. Tenzen, S. Giles, S. Zhong, W.J.R. Longabaugh, E.H. Davison, W.H. Wong and A.P. McMahon, submitted for publication), estrogen receptor (human) (2), Oct4, Sox2 and Nanog (human) (1), and ChIP-PET data for p53 (human) (3), Oct4 and Nanog (mouse) (4). The five ChIP-chip datasets were generated using different platforms. The Gli ChIP-chip was generated using a custom array produced by Agilent Technology. 50–150 kb regions surrounding promoters and 3'-untranslated regions (3'-UTRs) of a selected set of genes were surveyed by 60mer oligo probes at a density of one probe per 125 bp. The Oct4, Sox2 and Nanog ChIP-chip were based on Agilent promoter arrays which surveyed Oct4, Sox2 and Nanog ChIP-chip were based on Agilent promoter arrays which surveyed Oct4, Sox2 and Nanog (human) (1), and ChIP-PET data for p53 (human) (3), Oct4 and Nanog (mouse) (4). The five ChIP-chip datasets were generated using different platforms. The Gli ChIP-chip was generated using a custom array produced by Agilent Technology. 50–150 kb regions surrounding promoters and 3'-untranslated regions (3'-UTRs) of a selected set of genes were surveyed by 60mer oligo probes at a density of one probe per 125 bp. The Oct4, Sox2 and Nanog ChIP-chip were based on Agilent promoter arrays which surveyed 8 to +2 kb promoter regions of all human genes using 60mer probes with an estimated probe spacing of 280 bp. The ER ChIP-chip was produced using Affymetrix chromosome 21 and 22 tiling arrays where 25mer probes were tiled at a density of 1 probe per 35 bp. These data were summarized in Table 1 and were analyzed using a unified protocol as described below.

**General data analysis protocol**

We applied the ChIP-chip peak detection tool TileMap (15) to define potential protein-binding regions for the five ChIP-chip datasets (Gli, ER, Oct4, Sox2 and Nanog). Enriched sequence patterns in high-quality binding regions were then identified through de novo motif discovery (8,9). In order to identify the key motif that may mediate sequence-specific protein binding, we compared different motifs’ relative enrichment levels in high-quality binding regions versus control genomic regions. Then, the key motif’s relative enrichment levels were used to refine the cutoff for defining binding regions which in turn were subject to further analysis of GC-content, phylogenetic conservation and physical distribution.

For the three ChIP-PET datasets (p53, Oct4 and Nanog), our analysis followed the order of de novo motif discovery, key motif ascertainment, analysis of GC-content, conservation and distributional properties.

**Initial definition of binding regions**

For Gli, Oct4, Sox2 and Nanog ChIP-chip (on Agilent arrays), we applied TileMap (15) to compute a moving average (MA) statistic for each probe. Probes with the MA statistic three standard deviations away from the global mean were used to define potential binding regions, resulting in 65 initial regions in Gli, 1262 initial regions in Oct4, 1220 initial regions in Sox2 and 1842 initial regions in Nanog. For ER ChIP-chip (on Affymetrix arrays), we applied a hidden Markov model (HMM) using TileMap to detect binding regions. We detected 107 initial regions using a posterior probability cutoff value of 0.9. The rationale for choosing the algorithms and cutoffs is explained in Supplementary Data S1 and S2. For p53, Oct4 and Nanog ChIP-PET, all regions reported by the original authors were included in our subsequent analysis. For all datasets, the number of initial regions and the criteria used to define them are summarized in Table 2.

The genomic coordinates of all human regions were converted into coordinates based on NCBI build 35 (hg17). All mouse regions were converted into NCBI build 34 (mm6) coordinates. Repeat-masked sequences of these two assemblies were downloaded from the UCSC genome browser (http://genome.ucsc.edu) and were used for all subsequent sequence analyses.

**De novo motif discovery**

A subset of high-quality regions from each dataset were selected for de novo motif discovery (Table 2). For Gli, Oct4, Sox2 and Nanog ChIP-chip, the high-quality regions were defined as regions with at least one probe whose MA statistic is four standard deviations away from the global mean. This resulted in 30 Gli regions, 388 Oct4 regions, 477 Sox2 regions and 728 Nanog regions. All 107 initial regions were used for de novo motif discovery in the ER ChIP-chip dataset. For p53 ChIP-PET, high-quality regions were defined as 323 PET3+ regions (i.e. regions with ≥3 overlapping paired-end ditags) as suggested by Wei et al. (3). For Oct4 ChIP-PET, all 1052 regions were used. For Nanog ChIP-PET, since the identity of the Nanog motif has not yet been clearly established, we only included 602 PET3+ regions in the de novo motif discovery to assure the high quality of input sequences.

**De novo motif discovery** was performed by running a Gibbs motif sampler (8,9) (Supplementary Data S3) three times independently. Each time, 10 motifs were sampled simultaneously. An initial motif length (L = 9, 12, 15) was specified for all motifs at the beginning of the sampling, and the motif lengths were then adjusted during the sampling.
Table 2. Summary of initial, high-quality and final ChIP-binding regions

| ChIP-chip on Agilent arrays | Regions used for de novo motif discovery | Final regions |
|-----------------------------|----------------------------------------|--------------|
| Criteria | MA statistics > mean + 3 SD | MA statistics > mean + 4 SD | $r_1 > 2$, $r_2 > 2$, $r_3 > 2$ (no. of TFBS) > 0.25 |
| Gli-chip | 65 | 30 | (no. of binding regions) in a bin |
| Oct4-chip | 1262 | 388 | 600 |
| Sox2-chip | 1220 | 477 | 900 |
| Nanog-chip | 1842 | 728 | 600 |

ChIP-chip on Affymetrix arrays

| Criteria | HMM posterior probability > 0.9 | HMM posterior probability > 0.9 | $r_1 > 2$, $r_2 > 2$, $r_3 > 2$ (no. of TFBS) > 0.25 |
|-----------|-------------------------------|-------------------------------|---------------------|
| ER-chip | 107 | 107 | (no. of binding regions) in a bin |
| ChIP-PET | All regions reported in the original studies | p53: PET3+ Oct4: All Nanog: PET8+ | All regions reported in the original studies |
| p53-PET | 542 | 323 | 542 |
| Oct4-PET | 1052 | 1052 | 1052 |
| Nanog-PET | 2947 | 602 | 2947 |

Other cutoff values were also tried but did not change the conclusions drawn in the paper. Conserved TFBSs were defined as a TFBS that resides within the top 10% most conserved genomic regions. Conservation is evaluated through conservation scores, either a phastCons score (for human hg17) (18) or a score based on a window’s percent identity measure (for mouse mm6) (Supplementary Data S5).

Examination of motif’s relative enrichment levels

Three statistics, $r_1$, $r_2$ and $r_3$, were defined to characterize relative enrichment levels of a motif in ChIP-binding regions compared to control regions. Assume that $n_1B$ counts how many times a motif occurs in ChIP-binding regions, $n_2B$ is the total length of non-repeat sequences in binding regions, $n_1C$ counts how many times the motif occurs in control regions, and $n_2C$ is the total length of non-repeat sequences in control regions. We define $r_1 = (n_1B/n_2B)/(n_1C/n_2C)$ as the relative enrichment level of the motif. Similarly, let $n_{kB}$ ($k = B$ or C) count the number of phylogenetically conserved motif sites in specified genomic regions, and let $n_{kC}$ count phylogenetically conserved non-repeat base pairs in the regions. $r_2 = (n_{3B}/n_{4B})(n_{3C}/n_{4C})$ then defines the motif’s relative enrichment level in phylogenetically conserved ChIP-binding regions. Note that $n_{3B}/n_{1B}$ is the percentage of motif sites that are conserved and $n_{4B}/n_{1B}$ is the percentage of genomic sequences that are conserved. Finally, $r_3$, defined as $(n_{3B}/n_{2B})(n_{3C}/n_{2C})$, characterizes the relative enrichment level of phylogenetically conserved sites in general ChIP-binding regions (not necessarily conserved). Note that $r_3/r_2 = (n_{4B}/n_{2B})(n_{4C}/n_{2C})$ characterizes whether or not ChIP-binding regions tend to be more phylogenetically conserved than control regions.

Random genomic controls, design-based controls and matched genomic controls

To evaluate a motif’s relative enrichment level, three types of control regions were prepared. ‘Random genomic controls’ were regions randomly chosen from the genome. Each region was 2 kb in length and 5000 regions were chosen for each dataset. ‘Design-based Controls’ were regions randomly...
chosen from part of the genome that was surveyed in the original ChIP study. For Oct4, Sox2 and Nanog ChIP-chip, these were 5000 segments (each 2 kb long) randomly chosen from −8 to +2 kb promoter regions around transcription start sites (TSS). For Gli, all regions tiled in the custom array were used as the design-based control. For ER, we used human chromosomes 21 and 22. For p53, Oct4 and Nanog ChIP-PET, design-based controls were the same as the random genomic controls. ‘Matched genomic controls’ were control regions carefully chosen to match the physical distributions of ChIP-binding regions. To choose the matched controls, for each dataset, we first annotated binding regions by their closest RefSeq genes. We then computed the distance between the centers of ChIP-binding regions and the neighboring genes’ TSS. The center of a binding region is defined as its middle point, i.e. (region start + region end)/2. Next, we randomly chose genes from the RefSeq database. For each chosen gene, a 2 kb region was picked up so that the distance between the gene TSS and the region center followed the same empirical distribution of distances between ChIP regions and their closest genes. The number of matched control regions was chosen to be a multiple of the number of high-quality binding regions. This resulted in 5010 regions for Gli-chip, 5029 for ER-chip, 4845 for p53-PET, 5044 for Oct4-chip, 4770 for Sox2-chip, 5096 for Nanog-chip, 5260 for Oct4-PET and 4816 for Nanog-PET.

Refining the cutoff for defining binding regions

After the key motif was ascertained from the relative enrichment level \( r_1, r_2 \) and \( r_3 \), binding regions initially defined by TileMap were binned according to their raw ChIP signal strength (defined as MA or HMM statistics). The relative enrichment level \( r_1, r_2 \) and \( r_3 \) of the key motif were computed for each bin. In general, the enrichment levels decreased as the raw ChIP signals went down (Supplementary Figure S10). We chose a cutoff to define our final binding regions by simultaneously requiring \( r_1 > 2, r_2 > 2, r_3 > 2, \) and \([\text{the number of motif sites in a bin}] > 0.25 \times \text{(the number of binding regions in the bin)}\). This results in 30 Gli regions, 80 ER regions, 600 Oct4-chip regions, 900 Sox2-chip regions and 600 Nanog-chip regions (Table 2).

For p53-PET, Oct4-PET and Nanog-PET, all regions reported by the original authors were used as our final results, resulting in 542 p53-PET regions, 1052 Oct4-PET regions and 2947 Nanog-PET regions after conversion to human hg17 or mouse mm6 assembly (Table 2). These final regions were used for subsequent GC-content, conservation and peak clustering analysis.

Clustering of binding regions

To test if binding regions have a clustering tendency, we constructed a null distribution as follows. We first mapped the key transcription factor binding motif to the genome (or part of the genome on which the original ChIP study was performed); we then simulated binding regions by randomly choosing them from the mapped TFBSs. The chosen TFBSs were assumed to be the center of the simulated binding regions and their distances define the random peak-to-peak distances. The number of selected regions was set equal to the number of observed ChIP-binding regions. The distance between simulated regions was then used to construct the null, to which distance distribution of observed ChIP-binding regions was compared. Here, observed ChIP-binding regions are defined as binding regions that were obtained from the ChIP study and that contained at least one mapped TFBS of the key transcription factor. We used MATLAB gamfit and expfit functions to fit gamma and exponential distributions that describe the observed and simulated peak-to-peak distance, respectively. Maximum-likelihood estimates for the parameters are shown in Figure 4E–G.

RESULTS

Transcription factor binding motif can be unambiguously recovered by de novo motif discovery

In order to determine if current genome-wide ChIP technology allows unambiguous recovery of the DNA motif responsible for sequence-specific protein-binding, we applied Gibbs motif sampler (8,9) to the high-quality binding regions identified from the raw ChIP intensity data (see Materials and Methods). Gibbs motif sampler is a de novo motif discovery algorithm that searches for enriched sequence patterns in a collection of DNA sequences. The motifs are assumed to be unknown before the search. In all six cases where the genuine transcription factor binding motifs were indeed known from previous biological studies (Gli, ER, p53, Oct4-chip, Sox2-chip, Oct4-PET), the genuine motifs were successfully recovered (Figure 1 and Supplementary Figures S1–S8). In the p53-PET and Gli-chip studies, the genuine motif had the highest motif score among all the reported motifs. In the remaining cases, the genuine motif did not have the highest score, but when we compared the motif occurrence rate in ChIP-binding regions with the corresponding occurrence rate in matched control genomic regions (see below), the genuine motif stood out as the one with the highest relative enrichment level in all cases. Thus in all six cases where the transcription factor binding motif is indeed known, the motif can be unambiguously identified without any prior information on the motif itself. The recovery did not use information from cross-species conservation nor co-localization of motif sites. Both had been shown previously to be capable of increasing signal-to-noise ratio of motif discovery (10–14,19–24). Although it is possible that we may not be able to recover extremely weak motifs in future location studies, the results here at least gave us some confidence in applying the ChIP technology to identify previously unknown transcription factor binding motifs in mammalian genomes.

Certain sequence patterns tend to be detected in binding regions of multiple transcription factors

Certain motifs were found by Gibbs motif sampler in multiple unrelated datasets. The two most prominent examples are a GC-rich pattern GGGG[A/C/T]GGGG (denoted by G\(_n\)[H\(_n\)] thereafter) and a AT-rich sequence pattern TTTTTT (or T\(_n\))). The former was reported in all datasets except for ER-chip, whereas the latter was found ubiquitously (Figure 1).

Another example is a pattern CACACACA (or CA\(_n\)) which was found in p53-PET, Oct4-Sox2-Nanog-chip as
well as Oct4-Nanog-PET (Figure 1). All these ubiquitous motifs were found after repeat-masking. G_{\text{p}}H_{\text{p}} is similar to Sp1-binding pattern. Several previous studies implicitly suggested that this motif can be more easily found than other motifs. For example, Cawley et al. (25) performed ChIP-chip for Sp1, cMyc and p53. Although ‘the exact Sp1 consensus was recovered along with many slightly weaker variants’ by \textit{de novo} motif discovery, the recovery of cMyc

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1}
\caption{Comparisons of \textit{de novo} motif discovery from multiple ChIP studies. Eight experiments were examined here, including (A) Gli-chip, (B) ER-chip, (C) p53-PET, (D) Oct4-chip, (E) Sox2-chip, (F) Nanog-chip, (G) Oct4-PET and (H) Nanog-PET. For each factor, representative motifs recovered by \textit{de novo} discovery are shown. The motif score reported by Gibbs motif sampler for each motif is shown in parantheses. The complete lists of motifs reported by Gibbs motif sampler are present in Supplementary Figures S1–S8. For each reported motif, the relative enrichment level \( r_1 \), \( r_2 \) and \( r_3 \) were computed by comparing TFBS occurrence rates in ChIP regions to their counterparts in matched genomic controls (labeled by ‘Matched’) or random genomic controls (labeled by ‘Random’). The enrichment levels of all discovered motifs are compared here, and the data used to generate the figures are listed in Supplementary Tables S1–S8. The motifs responsible for the sequence-specific protein binding are underlined or indicated by arrows. For Nanog-chip and Nanog-PET, the motif responsible for the binding was unknown. In these two cases, Oct-Sox composite motif was highlighted, and the relative enrichment levels of the previously proposed Nanog motif (4) (labeled by ‘Nanog’) are also shown. For Gli-chip, ER-chip, Oct4-chip, Sox2-chip and Nanog-chip, enrichment levels in relative to design-based controls were also computed and are shown in Supplementary Figure S9.}
\end{figure}
and p53 motifs was not reported. When analyzing a set of muscle-specific regulatory regions, Zhou and Wong (12) reported that ‘all three algorithms (CisModule, BioProspector, MEME) successfully found the Sp-1 motif’, whereas the remaining four motifs cannot always be found. The fact that $G_nHG_n$ can be found ubiquitously raises an issue of how this motif should be interpreted when encountered in real studies. Although its frequent occurrence may reflect the general binding property of Sp1, it is also possible that the motif has roles other than Sp1 binding or the motif is part of heterogeneous sequence background that was not captured well by the sequence background models (usually a Markov model) used by various \textit{de novo} motif discovery algorithms (7–9,12–14,16). Other ubiquitous motifs such as $T_n$ and $CA_n$ are faced with a similar situation. It would be interesting for future studies to clarify the real roles of these ubiquitously occurring motifs. If they are part of sequence background, it would be useful to include them into a better background model to increase the sensitivity of \textit{de novo} motif discovery algorithms.

Besides the low-complexity patterns above, a non-trivial motif CCCAG occurs frequently too. This motif is part of the core Gli-binding consensus (Figure 2A). Indeed, when we restricted the \textit{de novo} motif discovery to the top 20 Gli-binding regions which included most high-affinity Gli-binding sites, the motif recovered was GACCA\textit{CCCAG}, a pattern consistent with the previously described Gli motif (26–28). The CCCAG pattern was also found in ER-chip (M2: CCCAGGCCCTG; M5: CCCAGCGCCTGGCC; M8: ATCCAGCTACTCT) (Figure 2D–F and Supplementary Figure S2), p53-PET (M8: CCCAGGGTCCCAG) (Figure 2C and Supplementary Figure S8). The consensus motif that induces Gli-specific protein binding is indeed GACCA\textit{CCCAG} (26,27). The motif GGGACTACAATTTCCCAGAA is enriched in both Sox2- and Nanog-binding regions identified from the ChIP-chip study (Figure 2B and Supplementary Tables S5 and S6). Several lines of evidence, including conservation, clustering and distribution in the proximity of the promoter, suggest that the motif is highly likely to be a functional promoter element (H. Ji and W.H. Wong, unpublished data). Other CCCAG-containing motifs also showed interesting patterns such as the repetitive occurrence of two half sites (ER-M2, ER-M5, p53-M8) (Figure 2C–E). The presence of this nucleotide module in TFBS for functionally disparate transcription factors raises the intriguing possibility that CCCAG can be used as a core for building different motifs. Alternatively, this sequence may provide some epigenetic cue that allows enhancer sites to become accessible, perhaps through chromatin remodeling.

\textbf{Properly chosen control regions are important for ascertaining key motifs}

Motifs non-specifically enriched in ChIP-binding regions may obscure the identification of the motif that is of main interest in each individual study. For example, the strongest motifs discovered by Gibbs motif sampler in Oct4-Sox2-Nanog-chip and Oct4-Nanog-PET were the ubiquitously occurring $G_nHG_n$ and $T_n$ sequences (Supplementary Figures S4–S8). Likewise, in ER-chip, a GC-rich pattern (M2) and the AT-rich pattern $T_n$ (M3) both had higher motif scores than the ER-binding element (Supplementary Figure S2). Many \textit{de novo} motif discovery methods [e.g. Gibbs motif sampler and MEME (7)] only rely on the use of positive sequences in which the motif is expected to be enriched. Recent studies suggest that negative sequences (i.e. genomic control regions where the motif is expected not to be enriched) may be used as an additional source of information to refine the motif discovery (6,29,30). Ideally, by comparing motif occurrence rate in positive sequences (i.e. ChIP-binding
regions) with its occurrence rate in negative sequences, one would hope that the key motif can stand out as the motif that has the highest relative enrichment level. To check if this is indeed the case in mammalian genomes, we did the following comparisons.

We first defined three statistical variables, $r_1$, $r_2$ and $r_3$, to characterize three different aspects of relative enrichment level (Materials and Methods). $r_1$ characterizes relative enrichment level of a motif in ChIP-binding regions as compared with control regions; $r_2$ characterizes relative enrichment level of phylogenetically conserved motif sites in phylogenetically conserved ChIP-binding regions; and $r_3$ characterizes relative enrichment level of phylogenetically conserved motif sites in general ChIP-binding regions (not necessarily conserved). We then prepared three sets of control regions for each dataset (Materials and Methods). ‘Random genomic controls’ are regions randomly chosen from the genome. ‘Design-based controls’ are regions randomly chosen from the part of the genome that was surveyed by the original ChIP study. ‘Matched genomic controls’ are control regions carefully chosen to match the physical distributions of ChIP binding, so that the distance between the simulated regions and the TSS of their neighboring genes has the same distribution as the distance between real binding regions and their neighboring TSS. When high-quality binding regions were compared to random genomic controls, the motifs of main interest stood out as the one with the highest relative enrichment level in some but not all datasets (Figure 1 and Supplementary Tables S1–S8). In Oct4-chip, Sox2-chip and Oct4-PET, for example, certain GC-rich motifs including G$_{(i)}$H$_{(j)}$ showed higher $r_1$ values than the Oct4-Sox2 composite binding motif. This suggests that the random genomic controls are not enough for controlling the compositional bias (e.g. GC-content) of ChIP-binding regions. In contrast, when matched genomic controls were used, in all six cases where the identity of the key motif was known (Gli, ER, p53, Oct4-chip, Sox2-chip, Oct4-PET), the motif of main interest stood out as the one with the highest $r_1$ (Figure 1 and Supplementary Tables S1–S8). The performance of design-based controls was intermediate between that of random controls and matched controls. Compared to random genomic controls, using design-based controls further resolved the key motifs in the promoter studies Oct4-chip and Sox-chip, and in the Gli-chip study where regions surveyed were biased towards promoters (Supplementary Figure S9). However, the key Oct-Sox motif in the genome-wide Oct4 ChIP-PET study was not ranked as the top motif, indicating that the design-based controls may still lack the ability to account for potential distributional bias in genome-wide studies (e.g. a bias where regions are more likely to be found near promoters). We note that in genome-wide studies, design-based controls and random genomic controls are equivalent. Together, our results suggest that random genomic controls and design-based controls may not be sufficient for controlling the compositional bias of ChIP-binding regions; therefore, carefully chosen matched control regions are important for discerning the key motifs from ChIP data. In future studies, this determination will be needed when the transcription factor binding motif is unknown and needs to be defined from the study itself. The comparisons here also alert ChIP data analysts to the fact that, while negative sequences can be used to improve various kinds of motif analyses, the analysis results could be affected greatly depending on how the negative sequences are chosen.

When we compared the key motif’s $r_1$, $r_2$ and $r_3$ in most cases (except for p53-PET), $r_2 > r_3$ (Figure 1), suggesting that ChIP regions tend to be more conserved than the genomic controls. In all cases using matched controls, $r_1 < r_2$ and $r_1< r_3$, suggesting that by combining cross-species information with sequence-specificity of TF binding, one can obtain higher signal-to-noise ratio in discriminating transcription factor binding sites with random noises.

Given what was observed in ER, Gli, p53, Oct4 and Sox2 data, we revisited the question of ‘what is the Nanog motif’. Until now, this motif has not been well established. By using matched genomic controls, the motif with the highest $r_1$, $r_2$ and $r_3$ in Nanog-chip data was the composite Oct-Sox motif (Figure 1F and Supplementary Table S6). The same motif also showed the highest $r_2$ and $r_3$ values in Nanog-PET data and had the second highest $r_1$ (Figure 1H and Supplementary Table S8). The motif with the highest $r_1$ in Nanog-PET was G$_{(i)}$H$_{(j)}$ whose $r_2$ and $r_3$ values were however lower. To summarize, in both mouse Nanog-PET and human Nanog-chip data, the strongest motif is the Oct-Sox motif. We also checked other motifs that had relatively high enrichment levels. After excluding the Oct-Sox motif, the two motifs with the highest $r_1$ value in Nanog-chip were a CCAAT-box (M10) and the CCCAG-containing motif GGGACTACAATTCCCAGAA (M15) which has been discussed in previous section (Supplementary Figure S6 and Supplementary Table S6). Both motifs were not found in Nanog-PET (Supplementary Figure S8) and are more likely to be motifs related to promoter regions instead of general Nanog binding, because the Nanog-chip data were based on promoter arrays. In Nanog-PET, no other motifs showed strikingly high relative enrichment levels other than the Oct motif M8 and a Sox2-like motif M9 (Supplementary Figure S8 and Supplementary Table S8). Loh et al. (4) reported a putative Nanog motif recovered from Nanog-PET. This motif was not discovered by our de novo motif discovery. When its relative enrichment level was checked, it showed the third highest $r_1$, $r_2$ and $r_3$ in Nanog-PET (Supplementary Table S8), but its enrichment level is much lower in Nanog-chip data (Supplementary Table S6). Loh et al. (4) noted that the motif cannot be consistently found by different motif discovery methods, and raised the possibility that ‘only a subset of the CATT-containing motifs is discovered’. Taken together, our data indicate that the primary DNA-binding region for Nanog is the Oct-Sox composite motif. The sequence analysis per se has not provided strong evidence to support the sequence-specific binding of Nanog via the recognition of the putative CATT-containing Nanog motif. This raises the possibility that Nanog does not directly recognize a sequence-specific motif, instead relying on association with Oct4 and/or Sox to bind DNA.

GC-content and cross-species conservation of ChIP-binding regions

One critical factor in all genome-wide ChIP studies, which seek to define a comprehensive cataloguing of target sites,
lies in how to define the threshold values for these sites. We defined the final ChIP-binding regions by choosing cutoffs using the key motif enrichment level (Materials and Methods). This resulted in 30 Gli regions, 80 ER regions, 600 Oct4-chip regions, 900 Sox2-chip regions, 600 Nanog-chip regions, 542 p53-PET regions, 1052 Oct4-PET regions and 2947 Nanog-PET regions. In all eight ChIP experiments, transcription factor binding regions had a higher GC-content than the genome-wide level (Figure 3A). This is consistent with the observation that random genomic controls are not enough to provide control for non-ChIP-specific GC-rich sequence patterns. Among the eight experiments, transcription factor binding regions had a higher GC-content than the genome-wide level (Figure 3A). This is consistent with the observation that random genomic controls are not enough to provide control for non-ChIP-specific GC-rich sequence patterns. Among the eight experiments, the high GC level in Oct4-Sox2-Nanog-chip and in Gli-chip can be explained by the experimental design which was either focused on or biased towards promoter regions. In the remaining studies, the relatively high GC-content in ChIP-binding regions cannot be explained by the design itself, reflecting the existence of a distributional bias of transcription factor binding (e.g. binding is more likely to happen near promoters). In all these cases, matched genomic control regions provided a comparative or better match to binding region GC-content than genome-wide controls or design-based controls.

In the five ChIP-chip datasets, the binding regions are more conserved phylogenetically than the genome-wide conservation level and the base-line conservation level determined by the experimental design (Figure 3B). On the contrary, regions identified from the three ChIP-PET experiments showed a much lower conservation level. This is consistent with Figure 1, where $r_f/r_2$ tends to be smaller in ChIP-PET studies than in ChIP-chip studies. Although this difference in conservation levels could be due to the difference in the design itself (i.e. promoter versus genome-wide studies) or due to the use of different transcription factors, one cannot completely rule out the possibility that the different conservation level here reflects the difference between the two technologies (see discussions in Supplementary Data S6). Which explanation is true should be resolved when more data from both technologies are coming available in future.
Binding regions of some transcription factors have a clustering tendency

We noticed that certain signature genes in each individual studies contained multiple binding regions identified by ChIP. For example, as a membrane receptor, Ptch1 plays a key role in transducing the sonic hedgehog (SHH) signal to Gli (31). The Gli-chip study identified five GLI-binding peaks within a 20 kb region of the Ptch1 promoter (see Figure 2A in Vokes et al., submitted for publication). Likewise, Nkx2-9 contained two peaks separated by 7 kb (Figure 2C in Vokes et al., submitted for publication). The known ER target, TFF1 (32), had at least two high-quality peaks (Figure 4A). Sox2, which is a key regulator in stem cell development (33–35) had two Oct4-binding regions and three Nanog-binding regions identified by ChIP-PET (Figure 4C). Jarid2, which was suggested to be an Oct4 target
in both human and mouse stem cells (1,4), contained at least two peaks for Oct4, Sox2 and Nanog binding (Figure 4B).

This observation led us to hypothesize that some transcription factor binding regions have certain clustering tendency. To check if this is indeed the case, for each of the three genome-wide or chromosome-wide datasets (ER-chip, Oct4-PET, p53-PET) and the Gli dataset where individual genes were covered substantially, we picked up binding regions that contained at least one occurrence of the key transcription factor binding motif. The distance between neighboring binding regions was computed and its distribution was compared to what was expected by random (Materials and Methods). Compared to the simulated random distance distributions, binding regions of Gli, ER and Oct4 showed a clear clustering tendency (Figure 4D–F). There is an elevated probability to observe a peak distance between 1 and 100 kb. For p53, this clustering tendency was not clear (Figure 4G).

Previously it was shown that 6–30 bp long transcription factor binding sites may be clustered within a cis-regulatory module 0.5–2 kb in length, and this type of clustering can be used to improve performance of motif discovery and CRM prediction (10–14). In contrast, our studies emphasize a higher level clustering of discrete cis-regulatory modules at a scale of 1–100 kb. The existence of this clustering tendency for certain transcription factors raises the question of whether this higher level clustering can be potentially used as an additional source of information to improve the prediction CRMs and target genes of transcription factors.

Detailed study on this issue is beyond the scope of our current paper, but a simple comparison in the Gli case provided in supplementary Data S7 is helpful to illustrate the potential. Paper, but a simple comparison in the Gli case provided in supplementary Data S7 is helpful to illustrate the potential. To check if this is indeed the case, for each of the three genome- or chromosome-wide datasets, we also tried to fit parametric models for the observed peak-to-peak distance distribution. In all three cases, the simulated distance distribution can be fitted well by an exponential distribution, suggesting that the random occurrence of binding regions can be characterized well by a Poisson process. On the other hand, the observed distance distributions can be fitted well only with a gamma distribution, which has much higher cumulative probabilities in the lower distance range.

Second, we showed that certain motifs occur ubiquitously in transcription factor binding regions. The real functions of these motifs need to be clarified by future biological experiments. Before this, we suggest that these simple sequence patterns should be interpreted with caution. In particular, for computational biologists, these ubiquitous motifs may not be a good choice for assessing sensitivity and specificity of de novo motif discovery algorithms especially when these algorithms are tested on mammalian ChIP-chip data. The result also emphasizes the need to develop better sequence generating models for modeling genomic background in motif discovery when these patterns are not of our main interest.

Third, the GC-content bias emphasizes the importance of choosing good genomic controls for ascertaining sequence patterns that are of main interest. Although several previous studies showed that negative sequences can help to improve the performance of motif discovery, the fact that the results can be affected greatly by the method used to choose negative sequences is less well-known.

Fourth, the clustering tendency of binding regions is a new piece of information that may be exploited in future to improve the computational prediction of cis-regulatory elements from mammalian genome. Although not used further in our current study, the observation that the peak-to-peak distance distributions follow a gamma distribution can be potentially incorporated into future statistical models for ChIP-chip data analysis (e.g. a model that reprioritizes peaks based on their clustering tendency). Besides its computational use, perhaps it is more interesting to understand the mechanisms behind the clustering of multiple binding regions. One possible explanation is that a gene may need multiple enhancers to confer tissue and context specific
response to a single transcription factor. Another possibility is that this clustering can raise the local concentration of TFBSs to a sufficiently high level so that the regulatory regions can be easily recognized by a transcription factor. The eventual clarification of the mechanisms warrants future investigation.

With these new observations, it is our hope that the comparative study here will facilitate a better understanding of the genome-wide location data for mammalian transcription factors, and help us to use future data more efficiently.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We thank the three anonymous reviewers for their helpful comments. This work is partially supported by NIH grant GM67250 (H.J. and W.H.W.) and the Helen Hay Whitney Foundation (S.A.V.). Funding to pay the Open Access publication charges for this article was provided by NIH.

Conflict of interest statement. None declared.

REFERENCES
1. Boyer,L.A., Lee,T.I., Cole,M.F., Johnstone,S.E., Levine,S.S., Zucker,J.P., Guenther,M.G., Kumar,R.M., Murray,H.L., Jenner,R.G. et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell, 122, 947–956.
2. Carroll,J.S., Liu,X.S., Brodsky,A.S., Li,W., Meyer,C.A., Szary,A.J., Eeckhoute,J., Shao,W., Hestermann,E.V., Geistlinger,T.R. et al. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell, 122, 33–43.
3. Wei,C.L., Wu,Q., Vega,V.B., Liu,X.S., Brodsky,A.S., Li,W., Meyer,C.A., Szary,A.J., Eeckhoute,J., Shao,W., Hestermann,E.V., Geistlinger,T.R. et al. (2005) Global map of p53 transcription-factor binding sites in the human genome. Cell, 124, 207–219.
4. Loh,Y.H., Wu,Q., Chew,J.L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. et al. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nature Genet., 38, 431–440.
5. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat. Biotechnol., 20, 835–839.
6. Hong,P., Liu,X.S., Zhou,Q., Xu,L., Liu,J.S. and Wong,W.H. (2005) A boosting approach for motif modeling using ChIP-chip data. Bioinformatics, 21, 2536–2643.
7. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, pp. 28–36.
8. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science, 262, 208–214.
9. Liu,J.S. (1994) The collapsed Gibbs sampler with applications to a gene regulation problem. JASA, 89, 958–966.
10. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. J. Mol. Biol., 278, 167–181.
11. Frith,M.C., Hansen,U. and Weng,Z. (2001) Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics, 17, 878–889.
12. Zhou,Q. and Wong,W.H. (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc. Natl Acad. Sci. USA, 101, 12114–12119.