Published online 18 May 2012

Nucleic Acids Research, 2012, Vol. 40, No. 17 ***e128***
doi:10.1093/nar/gks433

Inferring direct DNA binding from ChIP-seq

Timothy L. Bailey[1,*] and Philip Machanick[2]

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane 4072, Queensland, Australia and [2]Department of Computer Science, Rhodes University, Grahamstown 6140, South Africa

Received November 10, 2011; Revised April 2, 2012; Accepted April 23, 2012

Downloaded from https://academic.oup.com/nar/article/40/17/e128/2411117 by Rhodes University user on 27 May 2024

ABSTRACT

Genome-wide binding data from transcription factor ChIP-seq experiments is the best source of information for inferring the relative DNA-binding affinity of these proteins *in vivo*. However, standard motif enrichment analysis and motif discovery approaches sometimes fail to correctly identify the binding motif for the ChIP-ed factor. To overcome this problem, we propose 'central motif enrichment analysis' (CMEA), which is based on the observation that the positional distribution of binding sites matching the direct-binding motif tends to be unimodal, well centered and maximal in the precise center of the ChIP-seq peak regions. We describe a novel visualization and statistical analysis tool—CentriMo—that identifies the region of maximum central enrichment in a set of ChIP-seq peak regions and displays the positional distributions of predicted sites. Using CentriMo for motif enrichment analysis, we provide evidence that one transcription factor (Nanog) has different binding affinity *in vivo* than *in vitro*, that another binds DNA cooperatively (E2f1), and confirm the *in vivo* affinity of NFIC, rescuing a difficult ChIP-seq data set. In another data set, CentriMo strongly suggests that there is no evidence of direct DNA binding by the ChIP-ed factor (Smad1). CentriMo is now part of the MEME Suite software package available at http://meme.nbcr.net. All data and output files presented here are available at: http://research.imb.uq.edu.au/t.bailey/sd/Bailey2011a.

INTRODUCTION

Chromatin immunoprecipitation coupled with massively parallel sequencing (ChIP-seq) is a wonderful tool for studying the binding of transcription factors to genomic DNA. ChIP-seq provides a genome-wide map of the locations bound by the immunoprecipitated (ChIP-ed) transcription factor (TF). The resolution of the map depends on the TF and the software used to determine the binding locations (so-called 'peak-calling software'), but the predicted locations are often within 50 base pairs (bp) of a site matching the TF's known DNA-binding propensity (1). This map provides direct evidence of the enhancers and promoters bound by the TF and clues to its role in transcriptional regulation. In addition, the short genomic regions identified by ChIP-seq are generally very highly enriched with binding sites of the ChIP-ed TF, and consequently provide a rich source of information about its relative DNA-binding affinity. The regions also tend to be enriched for the binding sites of other TFs that bind cooperatively or competitively with the ChIP-ed TF (2,3).

DNA-binding motifs expressed as position-weight matrices (PWMs) can be used to model the binding free energy of a TF protein to a specific sequence of DNA relative to random DNA (4). (In what follows, we will simply say that a motif represents the 'DNA-binding affinity', dropping the term 'relative' for compactness of exposition.) A primary objective of many ChIP-seq experiments is determining the *in vivo* DNA-binding affinity of the ChIP-ed TF, and it has been shown that ChIP-seq tag densities are predictive of protein–DNA binding affinity (5). This is usually approached by *ab initio* motif discovery for which many algorithms exist (3,6,7). This approach results in one or more motifs, one of which may represent the DNA-binding affinity of the ChIP-ed TF. The other motifs may be those of cooperatively- or competitively-binding TFs. In many cases, one motif stands out as occurring more frequently in the ChIP-ed regions than any other, and is assumed to be that of the ChIP-ed TF.

Assuming that the most highly 'enriched' motif represents the direct DNA-binding affinity of the ChIP-ed TF can be dangerous for several reasons. Firstly, if the ChIP-seq data is of low quality due to poor antibody performance or sample preparation issues, the correct motif may not be present in the set of discovered motifs, or the algorithms may fail to find any motifs. Secondly, if the TF primarily binds DNA in conjunction with one or more other DNA-binding TFs, their motifs may appear more enriched than the ChIP-ed TFs. Thirdly, the ChIP-ed factor may not bind DNA directly at all, but always by 'piggy-backing' on one or more distinct DNA-binding TFs.

*To whom correspondence should be addressed. Tel: +61 7 3346 2614; Fax: +61 7 3346 2103; Email: t.bailey@imb.uq.edu.au

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article describes a novel method for identifying the DNA-binding motif of the ChIP-ed TF even in difficult ChIP-seq data sets. Our method is designed to overcome the first two sources of difficulty described in the preceding paragraph—poor ChIP-seq data quality or highly enriched co-factor binding sites. It can also predict when the third situation—binding by 'piggy-backing' is likely to be occurring. Our method can be used to analyze sets of motifs determined using *ab initio* motif discovery on the ChIP-seq regions. It can also be applied more generally as a motif enrichment analysis (MEA) tool (8–10), to consider all motifs in a compendium of known motifs as candidates for the ChIP-ed TFs binding motif.

Our analysis methodology, which we call 'central motif enrichment analysis' (CMEA), is based on the simple observation that the binding sites of the assayed transcription factor in a successful TF ChIP-seq experiment will cluster near the centers of the declared ChIP-seq peaks. In other words, the actual location of direct DNA binding by whatever protein or protein complex was actually pulled down by the antibody to the TF should *tend toward the center* of any given ChIP-seq region. This assumption should be true if the ChIP-seq region itself was identified based on sharply defined 'peaks' in the mapped sequence tag density, as is the case for many commonly used 'peak-calling' algorithms [e.g. MACS (11), PeakSeq (12), QuEST (13)]. When all goes well, the actual ChIP-ed binding site lies somewhere within a region of about 100 bp (1), centered on the 'peak', and with increasing probability closer to the center. In other words, we expect the probability (density) of the true binding location to be maximum in the center of a peak.

We implement our approach in the CentriMo algorithm (*Centrali*ty of *Mo*tifs), which takes as input a set of equal-sized regions identified in a TF ChIP-seq experiment and one or more TF binding motifs expressed as PWMs. Ideally, each of these ChIP-seq regions should be centered on a single coordinate reported as the position of 'maximum confidence' within a peak by the peak-calling software. If the program only reports regions (rather than single genomic positions), we use equal-sized genomic regions centered on the precise middle of each of the reported regions. For each motif, CentriMo outputs a plot of the probability that a predicted binding site occurs at each position in a ChIP-seq region (site-probability plot). It also outputs, for each motif, the width of the central region that is most enriched in binding sites according to a statistical test, and a *P*-value adjusted for multiple tests. We refer to this as the 'central enrichment *P*-value' of the motif. To aid in visualization, CentriMo outputs the site-probability curves for the *n* motifs that are most highly 'centrally enriched', according to their central enrichment *P*-values. Thus, CentriMo both serves as a visualization tool and provides an objective assessment of the degree to which each of the input motifs predicts centrally enriched binding sites.

As we show in the 'Results' section, CMEA is consistently able to determine the direct DNA-binding motif of ChIP-ed transcription factors, even in cases where motif discovery and motif enrichment algorithms fail or give ambiguous results. We illustrate how to apply CentriMo to analyze ChIP-seq data sets using motifs from motif discovery algorithms, motifs from motif databases and even hand-tailored motifs. In the process, we point out the characteristics of site-probability curves that distinguish between direct-binding motifs and the motifs of co-factors that merely bind near the ChIP-ed transcription factor with high frequency.

MATERIALS AND METHODS

The CentriMo algorithm

CentriMo begins by using a PWM motif to scan a set of equal-sized ChIP-seq peak regions. In each region, it declares at most one maximally scoring binding site, discarding any regions that have no match to the PWM above a given threshold. CentriMo resolves ties by randomly selecting one of the predicted binding sites with maximal score. We show below that the size of the PWM threshold is quite unimportant as long as it is above approximately three bits. CentriMo counts the number of sequences with a declared binding site that starts in each possible position in the peak regions, normalizes the counts to estimate probabilities and plots the resulting histogram after smoothing. CentriMo then efficiently computes the number of sequences with the declared binding site in each possible window centered on the middle of the ChIP-seq peak regions, and applies a one-tailed binomial test to the significance of any central enrichment of the declared sites.

Declaring at most one binding site in each equal-sized ChIP-seq peak region makes the statistical analysis extremely robust and simple. Since we discard regions with no declared site, each remaining region contains one site, and our null model assumes that the site is uniformly distributed within the region. This implies that a binomial model applies for the number of sites in any central window. Complications from overlapping sites do not arise since we only count one site in each region. Our approach solves the issue of how to choose background sequences or sequence models faced by other MEA approaches, since the flanking sequence around the central window serves this purpose. If the ChIP-seq peak region contains multiple identical actual binding sites, randomly choosing one maximal predicted site rather than discarding the region preserves much of the available information if the actual sites are near each other.

In more detail, the CentriMo algorithm is as follows. The primary input to CentriMo is a set of equal-length genomic sequences, each centered on a ChIP-seq peak. CentriMo predicts the best site in each sequence (and its reverse complement) of length L using a log likelihood ratio PWM motif, and counts the number of sequences where the best site occurs in position i, for $i = [1,\ldots, L]$, where the position is relative to the 5′-end of the sequence (or its reverse complement). Ties for best site within a sequence are broken randomly. Dividing these counts, c_i, by the number of sequences, n, gives an estimate of the probability distribution for the location of the best site. That is, the estimated probability that the best site in a

Downloaded from https://academic.oup.com/nar/article/40/17/e128/2411117 by Rhodes University user on 27 May 2024

given ChIP-seq peak region occurs at position i is $p_i = c_i/n$. CentriMo plots this distribution (the site-probability curve), shifting i so that the center of the plot is labeled as position zero. By default, CentriMo smooths the curve by averaging position bins of width 10. CentriMo also counts the number of sequences, S_w, that have their best site in the central w positions, and applies the binomial test to compute the 'central enrichment' P-value. Assuming that each position is equally likely *a priori* to contain the best site in a given sequence, the probability of the best site being in the central region is $P = w/(L-d+1)$, where d is the width of the motif. So, assuming the sequences are independent, the P-value is the probability of at least C_w best sites in the central region is given by the cumulative binomial distribution with parameters n trials, S_w successes and Bernoulli trial probability P. Central regions of all widths from 1 to $L/2$ are tested, so CentriMo adjusts the P-values for multiple tests using a Bonferroni correction.

We found that the above algorithm performs badly when many input sequences do not contain the motif. This will often be the case for co-factor motifs. It can also occur for the ChIP-ed TFs motif if the TF often binds DNA indirectly, or if the ChIP-seq peak data is of low quality for any number of reasons. Using the algorithm described above, CentriMo will not detect central enrichment, including in several of the ChIP-seq data sets used in this study.

A very simple change to the above algorithm, however, solves the problem. CentriMo simply discards any sequences that do not contain a match to the motif with log likelihood ratio score above a threshold. This means that CentriMo measures the central enrichment of predicted binding sites in sequences that appear to bind the TF represented by the motif. Intuitively, thresholding the PWM score removes sequences from consideration that are not likely to have been bound by the TF in question. By default, the threshold is quite low—five bits, corresponding to a likelihood ratio of 32. We found that the choice of score threshold is not critical, with values between three and eight bits yielding essentially identical results (see Supplementary Data). Of course, if the likelihood ratio threshold is higher than the maximum possible score for the given PWM, all sequences will be discarded. All results reported here use the default threshold (*score* $\geq$5 bits).

The input sequences (ChIP-seq peak regions) need to be long enough to include the bound sites and sufficient flanking region for the binomial enrichment test to be effective. ChIP-seq peaks typically will be within 50 bp of the binding event (1), so we expect the enriched central region to be approximately 100 bp wide. We conservatively choose 200 bp of flanking region on each side, and use 500 bp sequences as input to CentriMo in all results reported here. Users of CentriMo can easily judge from the shape of the site-probability plots if a different input sequence length is appropriate for their ChIP-seq data.

CentriMo does not adjust the enrichment P-values for the number of input motifs. This allows P-values to be compared among different runs that use the same ChIP-seq regions but different sets of input motifs. In practice, we find that the significance levels are so good that the best motifs would be extremely statistically significant even if their P-values were corrected for testing tens of thousands of motifs. Since this far exceeds the size of existing TF binding motif databases, we consider this extra level of multiple testing correction unnecessary.

CentriMo accepts position-frequency motifs in MEME format and converts them to log likelihood ratio PWMs (14). The background model for the PWM is the base frequencies in the input sequences. To transform a position-frequency motif to a PWM, we add 0.1 times the background frequency of the base to each cell of the position-frequency motif, divide by the background frequency, and compute the base-2 logarithm of the resulting likelihood ratio. Motif conversion utilities from other motif discovery algorithm formats and motif database formats are available in the MEME Suite software package (http://meme.nbcr.net/meme/doc/overview.html).

ChIP-seq data sets and motif databases

We use 13 mouse embryonic stem cell (ES) TF ChIP-seq data sets (15) and the mouse embryonic fibroblast (EF) ChIP-seq data set for NFIC (16). To prepare the Chen *et al*. (15) data sets for use we map the (centers of the) ChIP-seq peaks declared by the authors to the genome, and extract the 500 bp of genomic sequence centered on each peak in FASTA format. To prepare the NFIC data set for use we download the author-defined peaks from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo), `(GSM398010_NFI_peaks_wtMEF_300bp window_300bpvicinity_range.bed.gz)`, and use the UCSC genome browser to extract 500bp genomic regions centered 150 bp downstream of each given locus. (The given loci are 150 bp upstream of the centers of the declared peaks.)

For MEA, we use a compendium of motifs consisting of all vertebrate motifs in the JASPAR CORE database (17) plus all motifs derived for mouse TFs in the UniPROBE database (18). This compendium contains 532 motifs. We have made no effort to reduce the redundancy of the motif database because we believe doing so is generally unwarranted in central motif enrichment analysis. The statistical power of CMEA is extremely high, so the redundancy has little effect on its ability to detect enriched motifs, and duplicate motifs are often of varying (unknown) quality or were derived using different methods (e.g. *in vivo* versus *in vitro* methods), so including them in the analysis can be informative.

Hand-tailored motifs for Nanog

The consensus sequence of the motif found by DREME (3) in the Nanog ChIP-seq data, `CVATYA`, does not precisely match the motif reported from SELEX data, `MMATTA` (19), so we wondered if CentriMo could be used to decide which motif is more likely to be correct. The DREME motif's consensus agrees with the SELEX motif at three positions: `CVATYA`. At position 1, the DREME motif is more specific, allowing only `C`, whereas the SELEX motif allows either `A` or `C (M=A/C)`. Conversely, the DREME motif is more general at

Downloaded from https://academic.oup.com/nar/article/40/17/e128/2411117 by Rhodes University user on 27 May 2024

position 2 (V=A/C/G) than the SELEX motif (M=A/C), and at position 5 (Y=C/T). We run CentriMo on all eight possible combinations of the three variable positions in the two consensus motifs. We convert the consensus sequences to motifs using the iupac2meme program.

RESULTS

We analyze a total of 14 ChIP-seq data sets from mouse EF and ES cells. In the first data set, we show that strong central enrichment can be detected for the published *in vitro* motif for NFIC even though traditional motif discovery and motif enrichment algorithms might suggest that the ChIP-seq peaks are not enriched for NFIC binding sites. We then show that the binding affinity of Nanog in mouse ES cells appears to differ substantially from published *in vitro* data. This result supports previous evidence provided by motif discovery in the Nanog ChIP-seq data. In our third case study, we provide evidence that E2f1 binds DNA cooperatively with one or more transcription factors, especially YY1. Our analysis of the fourth ChIP-seq data set shows no evidence of direct binding for the ChIP-ed factor, Smad1. Finally, we apply CentriMo to 10 additional mouse ES cell ChIP-seq data sets to illustrate its general utility for inferring direct DNA binding in ChIP-seq data.

NFIC: agreement between *in vivo* and *in vitro* binding

Pjanic *et al*. (16) generated ChIP-seq data for NFIC, a member of the Nuclear Factor One (NFI) family of transcription factors, in mouse embryonic fibroblasts (EF). NFI family members are extremely important during development of mammalian neural and other tissues. Pjanic *et al*. (16) did not report using motif discovery on the ChIP-seq peaks, and utilized a binding motif based on *in vitro* binding data for NFIC (20,21) in analyzing their ChIP-seq data. Pjanic *et al*. (16) note that, because the PWM they use is based on *in vitro* binding data, it may not describe the binding specificity of NFIC *in vivo*.

We can confirm that it is difficult to determine the direct DNA-binding affinity of NFIC from this ChIP-seq data using approaches that do not consider central enrichment. We find that several *ab initio* motif discovery algorithms [Amadeus (6), DREME (3), MEME (22), Trawler (7) and WEEDER (23), see the Supplementary Data for details) applied to the NFIC (100 bp) ChIP-seq regions fail to discover a motif that matches the known *in vitro* motif, which has consensus TTGGCANNNTGCCAA. One algorithm, DREME (3), does find a motif with consensus sequence CHTGGC, which partially matches the NFI-C 'half-site' consensus TTGGCA, but this motif ranks 19th out of the 24 motifs that DREME discovers. (DREME ranks motifs in terms of the statistical significance of the number of ChIP-seq regions containing the motif compared with the number of shuffled versions of those sequences that contain the motif.)

When we apply a conventional MEA tool, AME (10), to the (500 bp) NFIC ChIP-seq regions, this motif (JASPAR ID MA0119.1), ranks 85th in terms of enrichment among the 532 motifs in the combined JASPAR and UniPROBE mouse motif databases. (AME ranks motifs by sorting the ChIP-seq regions and shuffled versions of them according to the total number of predicted binding sites they contain, and then applying the Fisher's exact test to calculate an enrichment *P*-value.) The overall (non-central) enrichment of the known motif for NFIC therefore appears to be quite low in this ChIP-seq data.

In contrast to the above results, the NFIC ChIP-seq regions are highly *centrally* enriched for the known, *in vitro* DNA-binding motif for NFIC (Figure 1a). Using CentriMo as a MEA tool, we see that the known NFIC motif shows much stronger central enrichment than any of the other 531 motifs in the combined JASPAR and UniPROBE motif databases. As seen in Figure 1a, which shows CentriMo results for the five most centrally enriched motifs, the distribution of the best predicted sites of the JASPAR NFIC motif in the ChIP-seq regions is much more centrally peaked than that of any other motif in the compendia. [We confirm that these conclusions are not affected by choice of PWM threshold (*score* $\geq$5 bits) in Supplementary Figure S5.) The maximum probability ($\sim$0.3%) occurs near the center of ChIP-seq regions, as we would expect if this motif represents the NFIC binding sites. The width of the region of maximum enrichment (w = 276) is quite large, suggesting that the resolution of the ChIP-seq peaks is not as good as the 50bp suggested in the literature for typical ChIP-seq experiments. The relative number of (500 bp) ChIP-seq regions containing the known motif (with *score* $\geq$5 bits) is also rather small (5307/39 807 = 13.3%). This could result from many causes—the motif is imperfect, NFIC often binds DNA indirectly, low ChIP antibody specificity or other experimental issues. At any rate, this example demonstrates the ability of central MEA to detect the presence of the binding motif of the ChIP-ed factor even in the presence of significant noise.

Three of the five motifs found by CentriMo to be most centrally enriched in the NFIC ChIP-seq regions show strong similarities to the known consensus TTGGCANNNTGCCAA (Figure 1b). In addition to the top-ranking JASPAR NFIC motif (MA0119.1), the JASPAR motif for the NFIC half-site (MA0161.1) ranks third, and the Hand1::Tcfe2a motif, which is similar to the TTGGCA half-site consensus, ranks fourth.

Nanog: *in vivo* binding differs from that predicted previously *in vitro*

The Nanog DNA-binding motif derived from SELEX experiments (19) is MMATTA (where M is either adenine or cytosine). We previously reported that none of the six motif discovery algorithms we applied to Nanog ChIP-seq data (15) discovers a motif matching this *in vitro* motif (3). In that work, we suggested that a similar motif discovered by DREME, with consensus CVATYA, might better describe the *in vivo* DNA-binding behavior of Nanog. Although CVATYA is similar to the *in vitro* consensus, it is far less enriched in the (100 bp) ChIP-seq regions (DREME $E = 10^{-64}$) than two other motifs discovered by DREME, which correspond to Sox2 and Oct4 (DREME $E = 10^{-243}$ and 10^{-101}, respectively).

Downloaded from https://academic.oup.com/nar/article/40/17/e128/2411117 by Rhodes University user on 27 May 2024

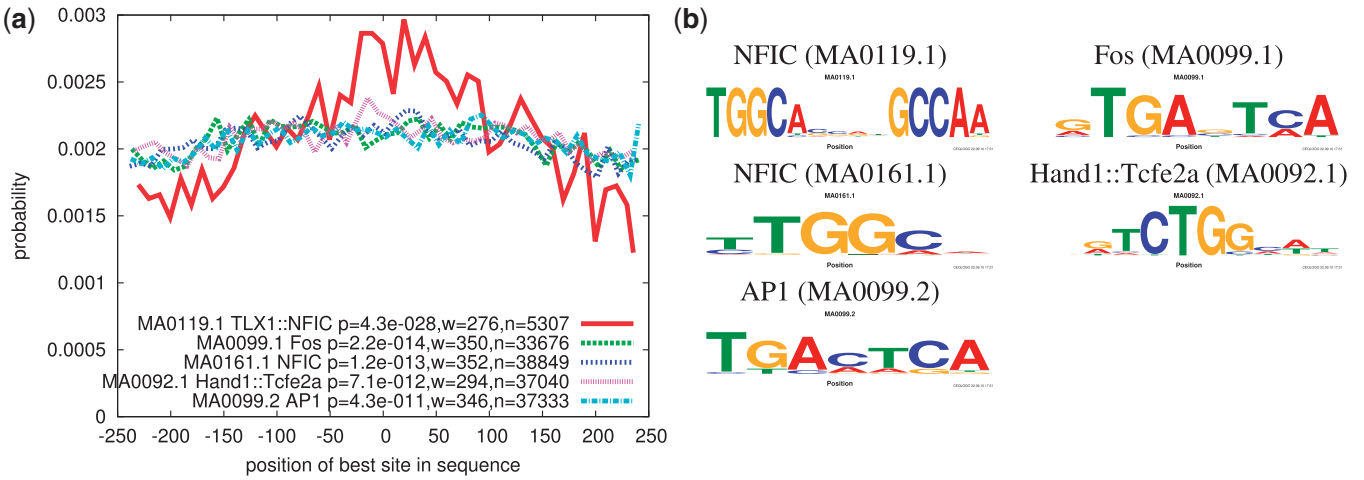

Figure 1. Confirming the *in vivo* DNA-binding affinity of NFIC. The top five CentriMo results for all JASPAR CORE and UniProbe mouse motifs (**a**) and their sequence logos (**b**) are shown. Each curve is the density (averaged over bins of width 10 bp) of the best strong site (*score* ≥5 bits) for the named motif at each position in the NFIC ChIP-seq (500 bp) peak regions from mouse EF cells. The legend shows the motif, its central enrichment *p*-value, the width of the most enriched central region (*w*), and the number of ChIP-seq regions (*n* out of 39807) that contain a motif site. JASPAR motifs MA0119.1 and MA0161.1 are known NFIC and NFIC half-site motifs, respectively.

So, based on these results, it is not clear that the CVATYA motif accurately describes Nanog's *in vivo* DNA-binding affinity.

A CentriMo analysis of the DREME motifs strongly suggests that the novel CVATYA is a good representation of the *in vivo* binding of Nanog in mouse ES cells (Figure 2a). Compared with the other motifs discovered by DREME in the Nanog ChIP-seq regions, CVATYA has a narrower region of maximum central enrichment ($w = 86$ bp), and it achieves a higher maximum site probability (~0.65%). Moreover, unlike the highly enriched Sox2 motif (ACAAWRS), the site-probability curve for CVATYA is unimodal and achieves its highest value precisely in the center of the ChIP-seq regions. The slight dips in the other site-probability curves near the center of the ChIP-seq regions suggest that those motifs represent binding by co-factors associated with Nanog. We note that, in contrast to the NFIC ChIP-seq data, the region of maximal central enrichment is much narrower for the CVATYA motif we propose for Nanog. In the Supplementary Data, we show that standard MEA does not recapitulate the central MEA results, and has the disadvantage of being quite sensitive to the length of the ChIP-seq peak regions used.

A comparison of levels of central enrichment strongly suggests that CVATYA better describes the *in vivo* binding of Nanog than MMATTA does (Figure 2**b**). The central enrichment of MMATTA is much less significant ($P = 2.5 \cdot 10^{-47}$ versus $P = 1.2 \cdot 10^{-208}$), much less narrow ($w = 134$bp versus $w = 86$), and achieves a much lower maximum site-probability than CVATYA. In fact, the six other similar motifs we selected for analysis all have more significant central enrichment than MMATTA. Three of these motifs—CMATYA, MMATYA and MVATYA—are also strong candidates for Nanog's *in vivo* binding motif. Each of these motifs, in position 5, allows a C as well as the T specified by the *in vitro* motif MMATTA. The MMATYA motif is particularly interesting because it differs from the *in vitro* motif only in position 5 and has maximum enrichment in a very narrow region ($w = 60$ bp). It seems clear from these results that Nanog DNA binding in mouse ES cells shows little preference for T over C in position 5, and is probably well-represented by any of the top four motifs in (Figure 2b).

The CVATYA motif strongly resembles another *in vivo* motif reported for Nanog (24) in mouse ES cells (Figure 3). This motif was found by motif discovery in the subset of Nanog ChIP-seq peaks that did not show binding by Sox2 or Oct4 in the Chen *et al.* (15) data. Although the central enrichment *P*-value of this motif is not as low ($P = 1.4 \cdot 10^{-164}$ versus $P = 1.2 \cdot 10^{-208}$), its region of maximal central enrichment is even narrower than that of CVATYA (44 bp versus 86 bp). This suggests that the He *et al.* (24) motif may be an even better description of *in vivo* binding by Nanog.

In subsequent work, the same group showed that a binding site predicted using the novel motif binds Nanog *in vitro*, but changing the adenine in the binding site (corresponding to position 6 in the CVATYA consensus motif and position 8 in the longer novel motif—see the inset in Figure 3) to either cytosine or thymine eliminates binding (25). This result agrees with our *in silico* analysis using CentriMo. When we modify the corresponding (last) position of the CVATYA motif to either C or T, the central enrichment drops dramatically (Figure 3, CVATYC and CVATYT). Thus, both *in vitro* data and central MEA of *in vivo* ChIP-seq data using CentriMo point to the importance of the adenine in position 6 of the CVATYA motif for Nanog binding.

E2f1: binding may be primarily indirect or cooperative

Previous studies of E2f1 binding in human (HeLa) cells indicated that few ChIP-ed E2f1 binding sites possess the canonical E2f1 motif (26). Central motif enrichment

Downloaded from https://academic.oup.com/nar/article/40/17/e128/2411117 by Rhodes University user on 27 May 2024

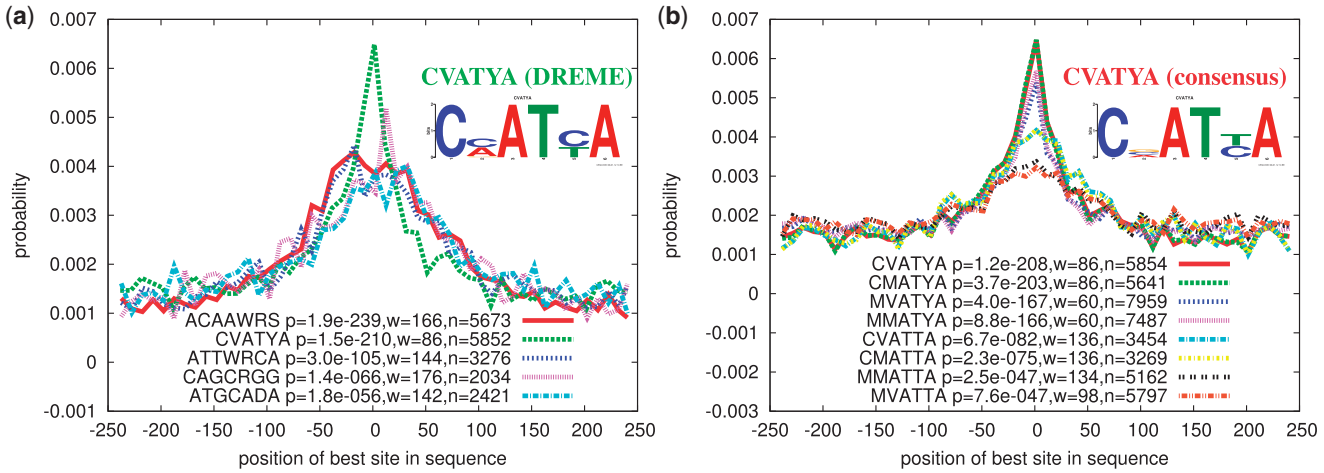

Figure 2. Inferring the DNA-binding affinity of Nanog in mouse ES cells. The top five CentriMo results for motifs discovered by DREME (**a**) and for consensus motifs similar to the SELEX-derived motif (**b**) are shown. Each curve shows the density (averaged over bins of width 10bp) of the best strong site (*score* ≥5 bits) for the named motif at each position in the Nanog (500bp) ChIP-seq peak regions. The legend shows the motif, its central enrichment *p*-value, the width of the most enriched central region (*w*), and the number of peaks (*n* out of 10343) that contain a motif site. The two CVATYA motifs (shown as sequence logos in the insets) differ slightly because the one in (a) is the PWM motif found by DREME, and the one in (b) is based solely on the consensus sequence.

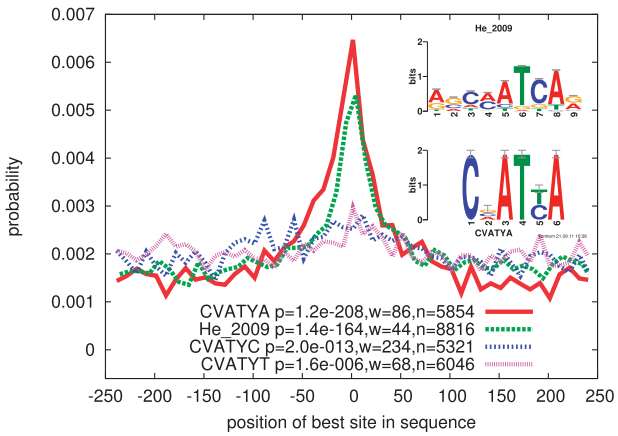

Figure 3. Central enrichment of two novel Nanog motifs (and two variants) in mouse ES cells. CentriMo results for the CVATYA motif discovered by DREME, the novel Nanog motif reported by He *et al.* (24), and two variants of the DREME motif are shown. The inset shows the aligned sequence logos of the two novel motifs. Each curve shows the density (averaged over bins of width 10 bp) of the best strong site (*score* ≥5 bits) for the named motif at each position in the (500 bp) Nanog ChIP-seq peak regions. The legend shows the motif, its central enrichment *p*-value, the width of the most enriched central region (*w*), and the number of peaks (*n* out of 10343) that contain a motif site. Note that the CVATYA motif used is based on the consensus sequence, not the PWM reported by DREME, and is the same motif as used in Figure 2b.

analysis of the Chen *et al.* (15) E2f1 ChIP-seq peak regions from mouse ES cells using CentriMo shows that none of the five E2f-family motifs contained in the JASPAR/UniPROBE compendium has a sharp central enrichment peak (Supplementary Figure S1). The most centrally enriched E2f-family motif, E2F3_secondary, has highly significant ($p = 4.2{\cdot}10^{-33}$) central enrichment, but the site-probability curve is very broad (Figure 4a). The lack of a narrow central enrichment peak with a large maximum site probability, as was observed for some Nanog motifs in the previous section, suggests that much of the binding of E2f1 to DNA is either indirect or in cooperation with another transcription factor. Either of these effects would modify the apparent binding motif of E2f1, and result in few ChIP-seq peaks containing the canonical binding site, in agreement with the previous results on E2f1 binding in human cells. In the Supplement (Supplementary Figure S6), we analyze E2f1 ChIP-seq data from HeLa cells, with essentially identical results, further supporting this hypothesis.

The most highly enriched motif in the E2f1 ChIP-seq peak regions in terms of both standard and central MEA is the JASPAR motif for GABPA (Figure 4a). However, like the motifs for the E2f-family members, the GABPA motif has a very broad site-probability curve. In fact, the curve is also bimodal, with maximal binding probability occurring approximately 50 bp on either side of the centers of the ChIP-seq peaks. This strongly suggests that indirect binding to DNA via GABPA is not responsible for the majority of the E2f1 ChIP-seq peaks, but, rather, with which GABPA often binds the genome in close proximity (within 50 bp) to E2f1.

In terms of the CentriMo central enrichment *P*-value, the maximum rank of any E2f-family motif is 14th out of 532 (E2F3_secondary). However, a different picture emerges when we sort the highly significant motifs—those with adjusted *P*-values no greater than 0.0001—by increasing 'width' of the region of maximal central enrichment. The top ranking motif is now YY1 (JASPAR MA0095.1), whose central enrichment region is far narrower ($w = 88$ bp) than those of the GABPA ($w = 268$ bp) and E2F3_secondary ($w = 232$bp) motifs (compare Figure 4a and b). Furthermore, the JASPAR motif for E2f1 now ranks second ($w = 128$bp), and its

Downloaded from https://academic.oup.com/nar/article/40/17/e128/2411117 by Rhodes University user on 27 May 2024

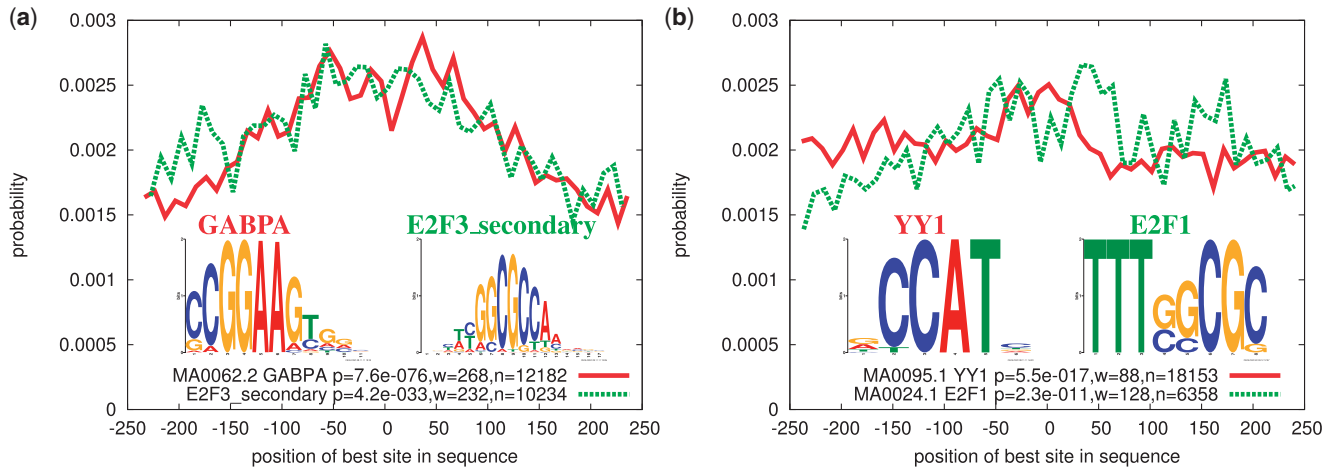

Figure 4. Central enrichment of E2f-family motifs and a other motifs in mouse ES cells. CentriMo results for the most enriched motif and the most enriched E2F-family motif in JASPAR/UniPROBE (**a**), and centrally enriched JASPAR motifs with narrow enrichment regions (**b**). Each curve in (**a**) shows the density (averaged over bins of width 10 bp) of the best strong site (*score* $\geq$5 bits) for the named motif at each position in the (500 bp) E2F1 ChIP-seq peak regions. The legend shows the motif, its central enrichment *p*-value, the width of the most enriched central region (*w*), and the number of peaks (*n* out of 20 699) that contain a motif site. YY1 and E2f1 rank first and second in terms of central enrichment among significant (P <0.05) JASPAR/UniPROBE motifs for which CentriMo predicts a central enrichment window narrower than 125 bp.

sequence logo and site-probability curve, along with those of YY1, are shown in Figure 4b. Although both of these motifs, especially YY1, show narrower central enrichment than that of the GABPA and other E2f-family member motifs, their maximum site probabilities are quite low (~0.25%). This casts doubt on whether either accurately describes the motif sites through which E2f1 may be indirectly or cooperatively binding the genome.

The fact that the YY1 motif has the narrowest region of maximal central enrichment among the 532 JASPAR/UniPROBE motifs is interesting given that YY1 has been suggested to interact with E2f-family members to stimulate transcription (27). That work, however, only demonstrated interaction between YY1 and E2f2 or E2f3, but not between YY1 and E2f1. Nonetheless, the CentriMo central enrichment results suggest that YY1 may interact with E2f1 and bind DNA cooperatively with it.

Smad1: CentriMo provides no evidence of direct binding in mouse ES cells

Previous studies (3,15) using motif discovery algorithms failed to discover a motif resembling the only known *in vitro* Smad-family motifs (Smad3_primary and Smad3_secondary from UniPROBE) in the Chen *et al.* (15) Smad1 ChIP-seq data set. Central enrichment analysis using all motifs in the JASPAR+UniPROBE compendium also does not suggest direct binding to sites matching either of the known Smad-family motifs. The top four CentriMo site-probability curves (Figure 5a) show evidence of binding to Oct-family—Pou5f1 (MA0142.1), Pou2f3_3986.2, Pou2f2_3748–and Sox-family—Sox2 (MA0143.1)—sites. On the other hand, there is a striking lack of evidence of central enrichment of either of the Smad3 motifs (Figure 5b), the only Smad-family motifs in the compendium. The central enrichment *P*-value for both of these Smad-family motifs is greater than 0.99. Clearly, predicted binding sites for this Smad-family motif are not enriched in the centers of the ChIP-seq peak regions from this experiment.

Non-central MEA suggests that there may be some binding to sites matching the UniPROBE Smad3_secondary motif, but not the Smad3_primary motif. AME gives Smad3_secondary a significant enrichment score ($P = 2 \cdot 10^{-8}$, data not shown). However, AME gives 27 of the 532 motifs in the JASPAR/UniPROBE compendium more significant scores than the UniPROBE Smad3_secondary motif. Both AME and CentriMo rank the same Oct4 motif (MA0142.1) first. These results suggest that Smad1 binds DNA primarily *indirectly* via Oct4 in mouse ES cells. It is also possible that the DNA binding of Smad1 differs substantially from that of Smad3, or that the ChIP-seq experiment was unsuccessful in locating binding of Smad1.

Central MEA is as effective as standard motif analysis

We have demonstrated the power of CMEA on three of mouse ES cell ChIP-seq data sets from Chen *et al.* (15)—Nanog, E2f1 and Smad1. We now show that central motif enrichment analysis is at least as effective as standard motif enrichment analysis on the remaining 10 ChIP-seq from that study (Figure 6). For standard enrichment analysis we use the AME algorithm, and for both types of enrichment analysis we use all motifs in the combined JASPAR/UniPROBE data set. In each of the 10 remaining sets of ChIP-seq regions, the motif with the most significant central enrichment *P*-value is the ChIP-ed factor's motif, a motif for a TF in the ChIP-ed factor's TF-family (e.g. Myc in the cMyc data set), or a motif for a heterodimer containing the TF (e.g. Oct/Sox heterodimer). The same is true of the top motif predicted by AME, except for the Zfx data set. In every case, for

Downloaded from https://academic.oup.com/nar/article/40/17/e128/2411117 by Rhodes University user on 27 May 2024

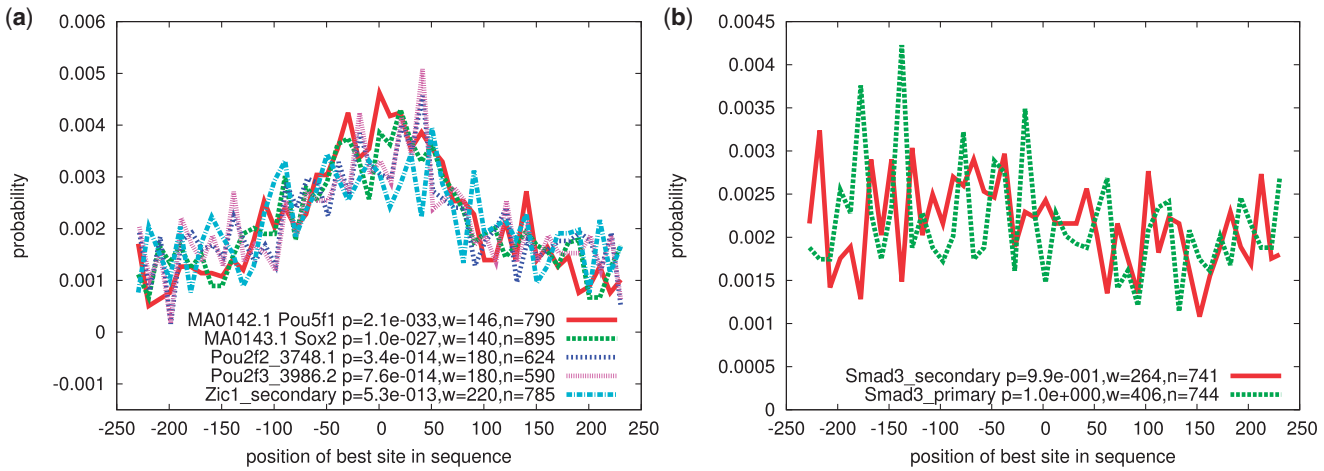

Figure 5. No evidence of direct binding by Smad1 in mouse ES cells. Results of CentriMo MEA analysis using the JASPAR+UniPROBE motifs (**a**) and using just the *in vitro* Smad3 motifs from UniPROBE (**b**) on the Chen *et al.* (15) ChIP-seq data. Each curve shows the density (averaged over bins of width 10 bp) of the best strong site (*score* ≥5 bits) for the named motif at each position in the (500 bp) Smad1 ChIP-seq peak regions. The legend shows the motif, its central enrichment *p*-value, the width of the most enriched central region (*w*), and the number of peaks (*n* out of 1126) that contain a motif site. JASPAR motifs MA0142.1 and MA0143.1 are for transcription factors Pou5f1 (Oct4) and Sox2, respectively.

both algorithms, the motif for the ChIP-ed factor is in the top two predicted motifs (Figure 6 caption).

Several of the plots of the site-probability curves in Figure 6 show very little variation among the top five motifs. This is because, for many of these 10 TFs, the JASPAR/UniPROBE data set contains several motifs for the TF, its family members or heterodimers. Nonetheless, the shape (width, height, centrality, unimodality) of the curves in all cases strongly suggests that the motifs correspond to the ChIP-ed factor or a partner via which it binds DNA, rather than to a non-interacting co-factor. The one exception is STAT3, which shows a very slight dip exactly in the center of the site-probability curve. Although not nearly as pronounced, this is reminiscent of the situation with Nanog (Figure 2**a**), and it remains possible that the *in vivo* STAT3 DNA-binding motif in mouse ES cells differs somewhat from that given in JASPAR.

We also test CMEA on seven c-Myc ChIP-seq data sets from human cell lines (HeLaS3, K562 and NB4) from the ENCODE project (28) and find that a motif for a Myc-family TF is in the top two based on the central enrichment *P*-value in all seven cases (Supplementary Figure S7). These combined results show that when the motif database used for CMEA contains a motif for the ChIP-ed TF or a member of its TF-family, that motif is highly likely to show the most statistically significant central enrichment.

DISCUSSION

Direct DNA binding by transcription factors can be inferred from the central enrichment of predicted binding sites in ChIP-seq peak regions. The CentriMo algorithm provides visual and statistical information that facilitate this inference. A combination of central enrichment features appears to be informative: (1) low *P*-value; (2) narrow region of optimal enrichment (<100 bp); (3) high maximal site-probability; (4) unimodal site-probability curve (e.g. not like ACAAWRS in Figure 2a). Using these features, we have extracted strong evidence about the *in vivo* DNA-binding affinity of NFIC, Nanog, E2f1 and Smad1 from ChIP-seq data that is not apparent without considering central enrichment. CentriMo is an important new tool for the biologist studying ChIP-seq data.

Central motif enrichment analysis can be applied to ChIP-seq data in several contexts. Firstly, it can be used to evaluate motifs reported by motif discovery algorithms to determine the most likely candidates for direct DNA binding in a ChIP-seq experiment. Secondly, it can be used as a motif enrichment tool in conjunction with databases of known motifs, especially in cases where motif discovery algorithms fail. (In this application, CentriMo has the advantage relative to other MEA approaches of not requiring selection of a set of control sequences, since the flanking regions provide the control.) Finally, it can be used to investigate the *in vivo* validity of *in vitro* motifs from the literature, as well as variations of those motifs, as we do here for Nanog. All of this suggests, of course, that central enrichment can be applied directly in a motif discovery algorithm, and we are currently developing such algorithms.

The CentriMo algorithm resembles the earlier 'position-analysis' algorithm (29,30). That algorithm detects words that show a heterogeneous distribution of occurrences within a set of input sequences. CentriMo generalizes this approach to predicting TF binding sites represented by a PWM (rather than individual words), and specializes it by confining the search to enrichment of a central region. 'Position-analysis' counts all occurrences of a given word in each evenly-spaced bin of a preselected size (typically 50 bp), and detects deviation from a uniform background using a Chi-squared test.

Downloaded from https://academic.oup.com/nar/article/40/17/e128/2411117 by Rhodes University user on 27 May 2024

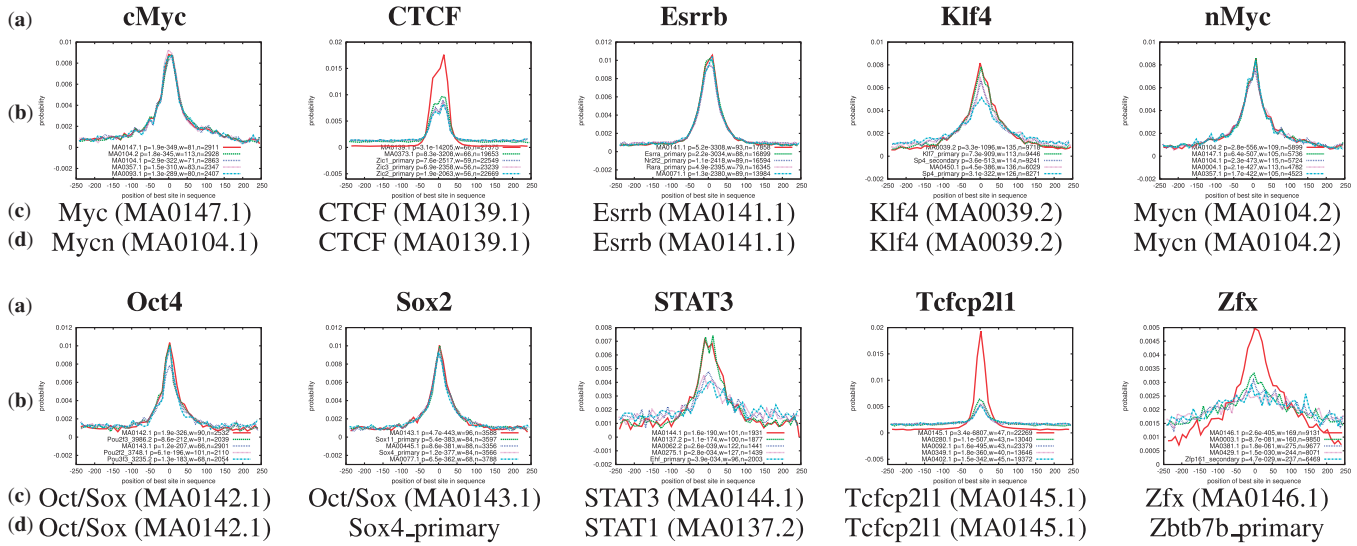

Figure 6. Comparison of CentriMo and AME for MEA of mouse ES cell ChIP-seq. Table shows results of applying CentriMo and AME to 10 ChIP-seq data sets from Chen *et al.* (15) using all 532 JASPAR CORE vertebrate and UniPROBE mouse motifs. Rows show: (**a**) the name of the ChIP-ed TF, (**b**) the CentriMo site-probability curves for the five most centrally enriched motifs, and the most enriched motif according to (**c**) CentriMo or, (**d**) AME. JASPAR motifs are given as TF name (JASPAR ID). We note that: the second-ranking CentriMo motif for Oct4 is Pou2f3_3986, an *in vitro* Oct motif; the second-ranking CentriMo motif for Sox2 is Sox11_primary, an *in vitro* Sox motif; the second-ranking AME motif for STAT3 is STAT3 (MA0144.1); and, the second-ranking AME motif for Zfx is Zfx (MA0146.1).

This contrasts with the approach of CentriMo, which allows each sequence to contribute at most one predicted binding site, making the statistical model significantly simpler. Unlike 'position-analysis', CentriMo does not require a bin size because it considers central regions of all size and computes the enrichment *P*-value for each using the binomial test.

CentriMo also bears some similarity to our recent SpaMo algorithm (2). The SpaMo algorithm detects enriched 'spacings' in ChIP-seq regions between predicted binding sites using *two* PWMs. In a typical application of SpaMo, one of the PWMs represents the DNA-binding affinity of the ChIP-ed TF, and the other a suspected co-factor. In each ChIP-seq region, SpaMo first predicts a binding site for the first PWM. It then aligns the regions on this binding site, and then predicts a binding site using the second PWM. SpaMo counts the number of times the predicted sites have each possible order, strand orientation and distance, and applies a binomial test to each of the combinations to detect significantly enriched spacings. CentriMo functions analogously, but the situation is simpler because the strand matched by the (single) PWM does not matter, nor does order (which side of the center a predicted site is located).

The distribution of distances between binding sites predicted by motifs and ChIP-seq peaks has been used previously to compare the quality of ChIP-seq peak-calling algorithms (31). In the current work, we use the distribution of distances for a completely distinct task—for comparing the quality of motifs as candidate descriptions of the direct DNA binding of the ChIP-ed TF. It is possible that Gerstein *et al.* (32) had something similar in mind with the 'localization tests' mentioned in the Supplement to that paper. However, they do not give details of their approach for computing the distance distributions or for computing their statistical significance, which is an essential facet of the analysis method we describe. To the best of our knowledge, the algorithm we describe here, and, especially, the associated methodology for inferring direct DNA binding, are both novel.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3, Supplementary Figure 1–7 and Supplementary References [33–39].

ACKNOWLEDGMENTS

We thank Ralf Jauch for suggesting we explore Nanog binding.

FUNDING

Funding for open access charge: National Institutes of Health [R0-1 RR021692-05 to T.L.B.].

Conflict of interest statement. None declared.

REFERENCES

1. Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, **5**, e11471.
2. Whitington,T., Frith,M.C., Johnson,J. and Bailey,T.L. (2011) Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res*, **39**, e98.
3. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.
4. Stormo,G.D. (1998) Information content and free energy in DNA–protein interactions. *J. Theor. Biol.*, **195**, 135–137.

Downloaded from https://academic.oup.com/nar/article/40/17/e128/2411117 by Rhodes University user on 27 May 2024

5. Jothi,R., Cuddapah,S., Barski,A., Cui,K. and Zhao,K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.
6. Linhart,C., Halperin,Y. and Shamir,R. (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res.*, **18**, 1180–1189.
7. Ettwiller,L., Paten,B., Ramialison,M., Birney,E. and Wittbrodt,J. (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat. Methods*, **4**, 563–565.
8. Frith,M.C., Fu,Y., Yu,L., Chen,J.-F., Hansen,U. and Weng,Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.*, **32**, 1372–1381.
9. Roider,H.G., Manke,T., O'Keeffe,S., Vingron,M. and Haas,S.A. (2009) PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, **25**, 435–442.
10. McLeay,R.C. and Bailey,T.L. (2010) Motif enrichment analysis: A unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, **11**, 165.
11. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
12. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnol.*, **27**, 66–75.
13. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M. and Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.
14. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
15. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
16. Pjanic,M., Pjanic,P., Schmid,C., Ambrosini,G., Gaussin,A., Plasari,G., Mazza,C., Bucher,P. and Mermod,N. (2011) Nuclear factor I revealed as family of promoter binding transcription activators. *BMC Genomics*, **12**, 181.
17. Sandelin,A., Alkema,W., Engström,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
18. Berger,M.F. and Bulyk,M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.*, **4**, 393–411.
19. Jauch,R., Ng,C.K.L., Saikatendu,K.S., Stevens,R.C. and Kolatkar,P.R. (2008) Crystal structure and DNA binding of the homeodomain of the stem cell transcription factor Nanog. *J. Mol. Biol.*, **376**, 758–770.
20. Roulet,E., Busso,S., Camargo,A.A., Simpson,A.J.G., Mermod,N. and Bucher,P. (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol*, **20**, 831–835.
21. Osada,S., Daimon,S., Nishihara,T. and Imagawa,M. (1996) Identification of DNA binding-site preferences for nuclear factor I-A. *FEBS Lett.*, **390**, 44–46.
22. Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, Vol. 3, Cambridge, UK. July 16–19, 1995. pp. 21–29.
23. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.
24. He,X., Chen,C.-C., Hong,F., Fang,F., Sinha,S., Ng,H.-H. and Zhong,S. (2009) A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS One*, **4**, e8155.
25. Xie,D., Chen,C.-C., Ptaszek,L.M., Xiao,S., Cao,X., Fang,F., Ng,H.H., Lewin,H.A., Cowan,C. and Zhong,S. (2010) Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res.*, **20**, 804–815.
26. Bieda,M., Xu,X., Singer,M.A., Green,R. and Farnham,P.J. (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res.*, **16**, 595–605.
27. Schlisio,S., Halperin,T., Vidal,M. and Nevins,J. (2002) Interaction of YY1 with E2Fs, mediated by RYBP, provides a mechanism for specificity of E2F function. *EMBO J.*, **21**, 5775–5786.
28. ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*, **9**, e1001046.
29. van Helden,J., del Olmo,M. and Pérez-Ortín,J.E. (2000) Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.*, **28**, 1000–1010.
30. Thomas-Chollier,M., Sand,O., Turatsinze,J.-V., Janky,R., Defrance,M., Vervisch,E., Brohée,S. and vanHelden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–W127.
31. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol.*, **26**, 1351–1359.
32. Gerstein,M.B., Lu,Z.J., Nostrand,E.L.V., Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K. *et al.* (2010) Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, **330**, 1775–1787.

Downloaded from https://academic.oup.com/nar/article/40/17/e128/2411117 by Rhodes University user on 27 May 2024