Published online 20 May 2011 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* ***W391–W399***
doi:10.1093/nar/gkr387

CENTDIST: discovery of co-associated factors by motif distribution

Zhizhuo Zhang[1], Cheng Wei Chang[1,2,3], Wan Ling Goh[2], Wing-Kin Sung[1,3,*] and Edwin Cheung[2,4,5,*]

[1]School of Computing, National University of Singapore, Computing 1, 13 Computing Drive, Singapore 117417, [2]Cancer Biology and Pharmacology, [3]Computational and Mathematical Biology, Genome Institute of Singapore, 60 Biopolis Street, #02-01 Genome, Singapore 138672, [4]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597 and [5]School of Biological Sciences, Nanyang Technological University, Singapore 637551

Received January 31, 2011; Revised April 28, 2011; Accepted May 3, 2011

ABSTRACT

Transcription factors (TFs) do not function alone but work together with other TFs (called co-TFs) in a combinatorial fashion to precisely control the transcription of target genes. Mining co-TFs is thus important to understand the mechanism of transcriptional regulation. Although existing methods can identify co-TFs, their accuracy depends heavily on the chosen background model and other parameters such as the enrichment window size and the PWM score cut-off. In this study, we have developed a novel web-based co-motif scanning program called CENTDIST (http://compbio.ddns.comp.nus.edu.sg/~chipseq/centdist/). In comparison to current co-motif scanning programs, CENTDIST does not require the input of any user-specific parameters and background information. Instead, CENTDIST automatically determines the best set of parameters and ranks co-TF motifs based on their distribution around ChIP-seq peaks. We tested CENTDIST on 14 ChIP-seq data sets and found CENTDIST is more accurate than existing methods. In particular, we applied CENTDIST on an Androgen Receptor (AR) ChIP-seq data set from a prostate cancer cell line and correctly predicted all known co-TFs (eight TFs) of AR in the top 20 hits as well as discovering AP4 as a novel co-TF of AR (which was missed by existing methods). Taken together, CENTDIST, which exploits the imbalanced nature of co-TF binding, is a user-friendly, parameter-less and powerful predictive web-based program for understanding the mechanism of transcriptional co-regulation.

INTRODUCTION

In order to precisely regulate the expression of target genes, transcription factors (TFs) bind to specific short stretches of DNA sequences or motifs in our genome. Generally, a gene is not regulated by only a single TF, but instead by a combination of TFs binding to chromatin in close proximity. For example, the Androgen Receptor (AR) and the forkhead factor, FoxA1, are co-localized together at AR-binding sites (ARBS) to regulate the transcription of AR-dependent genes in prostate cancer cells (1), whereas, Sox2, Oct4 and Nanog all converge together at enhanceosomes to control genes involved in pluripotency and self-renewal in embryonic stem (ES) cells (2). TFs that co-localize and collaborate together are known as co-associated TFs (or co-TFs) of each other.

Identifying co-TFs is an important step in understanding the mechanism of transcriptional regulation. Recent advances in ChIP-seq and the wide adoption of the technology in mapping TF-binding sites has allowed researchers to identify novel co-TFs (3). Currently, co-TFs of a selected TF are identified in the following manner. First, a peak calling program such as MACS (4) or CCAT (5) is used to determine which peaks in the ChIP-seq data are binding sites. Next, candidate co-TFs are predicted by examining if their motifs (position weight matrix, PWM) are enriched near the ChIP-seq peaks after normalizing against a chosen background model. TFs with enriched motifs are classified as potential co-TF candidates and subsequently validated experimentally. This approach,

*To whom correspondence should be addressed. Tel: +65-651-63580; Fax: +65-6779-4580; Email: ksung@comp.nus.edu.sg
Correspondence may also be addressed to Edwin Cheung. Tel: +65-6808-8184; Fax: +65-6808-8305; Email: cheungcwe@gis.a-star.edu.sg

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

known as the enrichment based method, has been widely used to identify novel co-TFs in web-based programs such as CEAS (6), CORE_TF (7), ConTra (8) and oPOSSUM (9). However, there are occasions when this approach fails to find co-TFs. This is because the accuracy of enrichment-based methods is highly dependent on several user-specific parameters including: (i) the background (which models the non-binding sites); (ii) the enrichment window size (which models the distance between the co-TF and the peak); and (iii) the PWM score (10) cut-off (which determines if a site can be bound by the co-TF or not). Since different co-TFs require different parameters, existing methods can only identify co-TFs that satisfy the parameters specified by the user. This restriction thus limits the accuracy of existing methods. To avoid this problem, it would be ideal to have a method that automatically determines the background and estimates the enrichment window size as well as the PWM score cut-off for every co-TF.

Recently, several studies showed that if two TFs are co-associated, their ChIP-seq peaks (or their binding sites) are not only in close proximity with each other, but the relative distance of each TF with respect to the other exhibits a peak-like distribution (1,2,11). We call this property the center distribution. Herein, we examine whether center distribution can be utilized for co-TF discovery. Moreover, we have developed a method called CENTDIST (http://compbio.ddns.comp.nus.edu.sg/~chipseq/centdist/), which ranks TFs based on their center distribution score. Unlike existing enrichment based methods, CENTDIST does not require any user-specific parameters. It can automatically optimize the enrichment window size and the PWM score cut-off. Furthermore, CENTDIST can predict weakly or marginally enriched co-TFs. In term of usability, CENTDIST is fast, user-friendly, and capable of handling data sets with over a million ChIP-seq peaks. The web-interface of CENTDIST also provides useful additional information that helps users select the best co-TF candidates.

We compared the performance of CENTDIST against two enrichment-based programs on 13 ChIP-seq data sets generated for 13 TFs from mouse ES cells (2). Our large-scale comparison showed that CENTDIST was the best performer amongst the three programs. We also applied CENTDIST on an AR ChIP-seq data set generated from a prostate cancer cell line. CENTDIST was sensitive enough to discover all known co-TFs (eight co-TFs) of AR within top 20 hits. Furthermore, CENTDIST identified AP4 as a novel co-TF of AR, which was not found by traditional enrichment-based methods. Taken together, CENTDIST is a powerful and user-friendly tool for studying the mechanism of TF co-regulation.

METHODOLOGY AND RESULTS

Imbalanced distribution of TF motifs around ChIP-seq peaks

Accurately predicting all the co-TFs of a particular TF from a ChIP-seq experiment is computationally challenging because some co-TFs may occur infrequently while the location of others are less certain than that of the ChIPed TF (Chromatin Immuno-precipitated TF in ChIP-seq experiment). Previous reports suggested that motifs of co-TFs are enriched around ChIP-seq peaks after normalizing against a particular background model (12,13). However, it is often difficult to choose the correct background model. Furthermore, it is also not easy to select the correct PWM score cut-off or the proper enrichment window size of co-TFs.

Instead of asking users to specify these parameters, we have developed a new program called CENTDIST which requires minimal input from users. Users only need to enter a set of genomic locations representing ChIP-seq peaks (chromosome-peak summit position) and a list of candidate PWM motifs [provided by users or obtained from either the TRANSFAC (14) or JASPAR (15) databases] representing co-TF-binding sites. Based only these two pieces of information, CENTDIST will compute the distribution of motif occurrences with respect to the peak summit (frequency graph) for each PWM motif under different PWM score cut-offs. CENTDIST will then find an optimal set of parameters that maximizes the frequency score (see below) near the ChIP-seq peaks. Given a frequency graph, and an enrichment window size d, we denote m_i and m_o to be the frequencies within and outside the enrichment window, respectively. The frequency score is defined as:

$$Z_{\text{Frequency}} = Z\left(m_i, \frac{m_o}{(m_i+m_o)}, m_i+m_o\right)$$

where $Z(x,\ p,\ n) = (x-np)/sqrt(np(1-p))$ is the normal approximation of the binomial Z-score for observing x successes out of n trials where the probability of a success trial is p. For example, we examined an AR ChIP-seq data set that was generated in our lab from the prostate cancer cell line, LNCaP. As shown in Figure 1a, the AR motif (RGAACANNNTGTTCY) occurs much more frequent near the center of the AR peaks, when compared to the flanking regions. Thus, the AR motif is imbalancely distributed and would be considered as having a good frequency score.

However, there are occasions when noise (like CG/AT bias) could also be imbalancely distributed around ChIP-seq peaks. Although such noise may be enriched, we expect it will not change dramatically near the center of ChIP-seq peaks compared to flanking regions. Therefore, to account for noise in the data, we included a function called the velocity score. The velocity score is derived from a velocity graph of the co-TF motif (Figure 1b) which is generated from the slope of the frequency graph (Figure 1a). If noise is assumed to change slowly (or linearly), the velocity of noise will be near zero; otherwise, it will change dramatically near to the peaks as compared to the flanking regions. Specifically, the velocity score is a Z-score which measures if the velocity is changed dramatically. Similar to frequency score, given a velocity graph and an enrichment window size d, we denote the positive and negative velocity within the

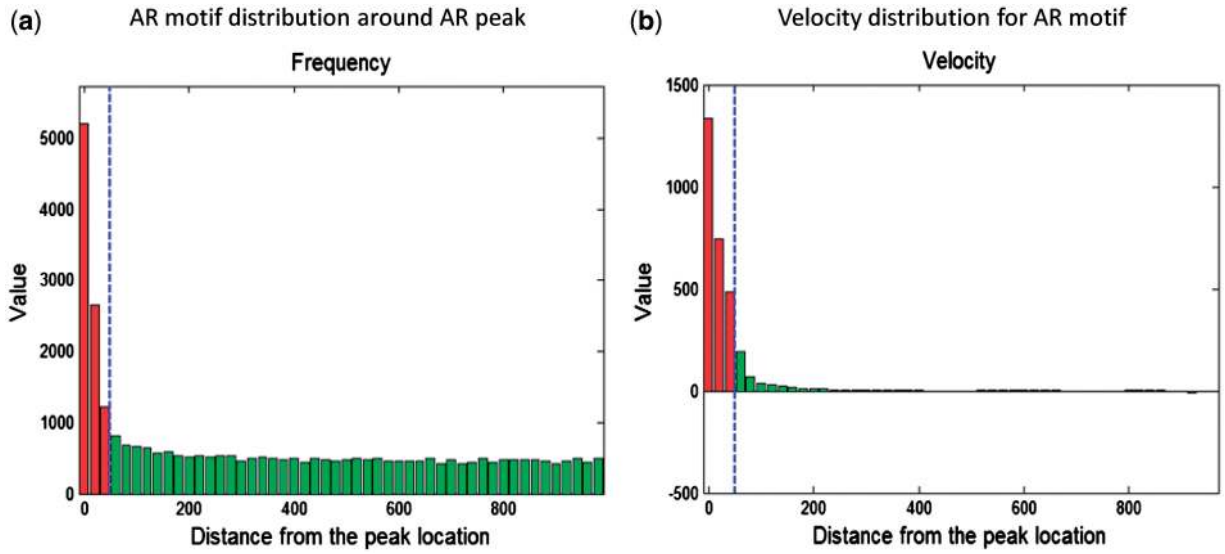

Figure 1. AR motif distribution analysis of our AR ChIP-seq data set. (**a**) Frequency graph of the AR motif in an AR ChIP-seq data set. (**b**) Velocity graph of the AR motif in an AR ChIP-seq data set. In each graph, the dotted line partitions the distribution into the enriched region (left region) and the flanking region. The generation of these two graphs can be found in Supplementary Section 1.1.

window m_{i+}, m_{i-} and outside the window m_{o+}, m_{o-} and the velocity score is defined as:

$$Z_{\text{Velocity}} = Z\left(|m_{i+}|+|m_{o-}|, \frac{(|m_{i-}|+|m_{o+}|)}{M}, M\right)$$

where $M=(|m_{i+}|+|m_{o-}|+|m_{i+}|+|m_{o-}|)$.

In short, CENTDIST will also take into consideration the velocity distribution of motif occurrences (velocity graph), which will correct the frequency score biases due to CG (or AT) variation in the regions around the ChIP-seq peaks. The scoring function used by CENTDIST to assess motif distribution is called the center distribution score (implementation details can be found in Supplementary Section 1), which is the sum of two components: frequency score and velocity score. For example, we observed a dramatic change in velocity (or slope) for the AR motif in the enriched region of the AR ChIP-seq peaks while the velocity remained constant in the flanking region (Figure 1b). In such instance, the AR motif would be classified as having a good velocity score.

Figure 2 demonstrates the capability of CENTDIST to promote true positive and repress false positive. To demonstrate the former, we consider the motif occurrence of V$AR_02 around AR ChIP-seq peaks. As shown in Figure 2a, the *Z*-score progressively increases as we use flanking region as background (instead of promoter or random region), select the optimal window, select the optimal PWM cut-off and finally considering the velocity. To demonstrate the latter, we study the CG-rich yeast TF motif, F$ADR1_01, which would have been determined incorrectly to be enriched around the Pol II (RNA polymerase II) ChIP-seq peaks in human K562 cells (16) using traditional approach. We know this motif is not actually enriched because Pol II-binding sites are enriched for CpG islands, which are regions known to contain many CG repeats. As shown in Figure 2b, this motif has a modest center distribution score based on only the frequency score, but the final center distribution score was significantly lower after taking the velocity score into consideration.

Verification of CENTDIST on a large scale ChIP-seq experiment

To determine if CENTDIST can identify co-TFs better than existing enrichment based methods, we compared the performance of CENTDIST with two motif scanners, CORE_TF (http://grenada.lumc.nl/HumaneGenetica/CORE_TF/) and CEAS (http://liulab.dfci.harvard.edu/CEAS/) (6,7). We chose CORE_TF and CEAS for our comparisons because they were the only programs we could find that report enriched motifs from user-defined genomic regions while other programs were limited to promoter regions only. For our comparisons, each program was optimized to their best performance (Supplementary Section 2.4).

Recently, 13 TF ChIP-seq maps were generated from mouse ES cells (2). These 13 TFs were shown to cluster into two core transcriptional modules called MTLs (multiple TF-binding loci). Because numerous co-TF relationships were discovered from the 13 factors, we decided to use these data sets for our comparisons of the three motif scanners. Only genomic locations of the ChIP-seq peaks and motifs from the TRANSFAC database were entered into CENTDIST. For CORE_TF and CEAS, input sequences with different window size (± 100, ± 200 and ± 500 bp) around the summit of the ChIP-seq peaks were extracted and different background settings were tested. The results from each program were compared against a table containing the co-TF motifs for each of the 13 ES TFs (Supplementary Table 7).

We assessed the performance of each program by the area under the receiver operating characteristic (ROC)

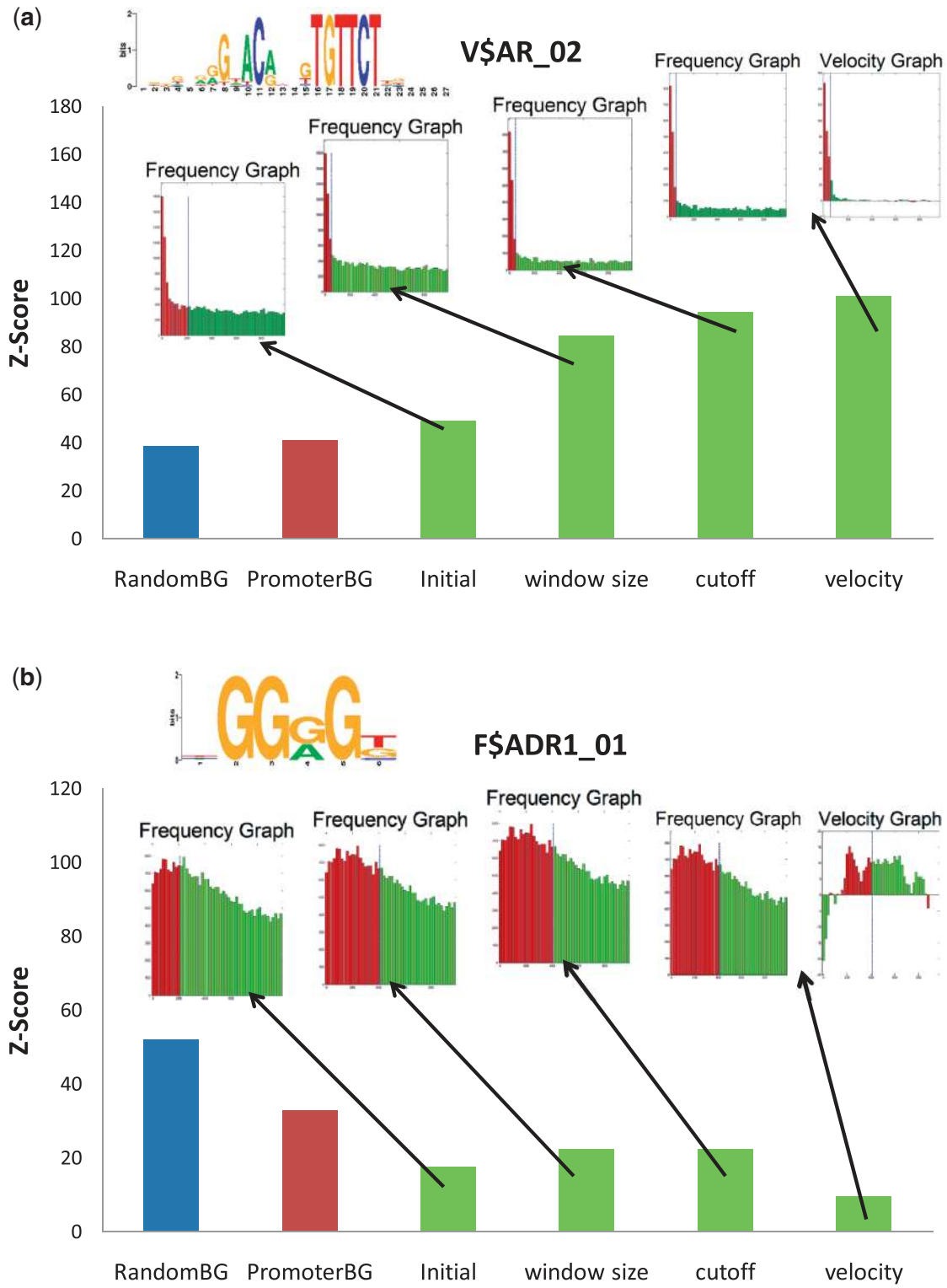

Figure 2. Demonstration of CENTDIST Capability. (**a**) CENTDIST enhances the *Z*-score of the AR motif in the AR ChIP-seq data set (LNCaP cell line). The blue bar and red bar show the *Z*-scores of the AR motif computed using the traditional enrichment method under the window size of 200 bp and the default PWM cut-off (1.32), respectively. The green bars show the *Z*-score of the AR motif computed by CENTDIST after it optimized different parameters. In the initial stage, the frequency *Z*-score was calculated using flanking regions at 200 bp as background and default PWM cut-off. In the second stage (window size), CENTDIST finds the best window size to maximize the *Z*-Score, in which the enrichment window size of AR is changed from 200 to 40 bp. In the third stage (cut-off), CENTDIST finds the best PWM cut-off to maximize the *Z*-Score, which leads to the flanking region noise level dropping significantly. In the fourth stage, CENTDIST combines the *Z*-scores of both the frequency graph and the velocity graph, thus further increasing the *Z*-Score. (**b**) CENTDIST can repress the *Z*-score of the false CG-rich motif in the Pol II ChIP-seq data set compared to the traditional overrepresentation methods. All *Z*-scores are computed exactly as in (a). Since CENTDIST considers the velocity graph of the false CG-rich motif, the combined *Z*-score of CENTDIST finally drops and is significantly lower than that computed by the traditional enrichment based method. As a side note, this figure also showed that random background can produce quite different results compared to promoter background, which highlights the difficulty of choosing a correct background in existing enrichment based methods.

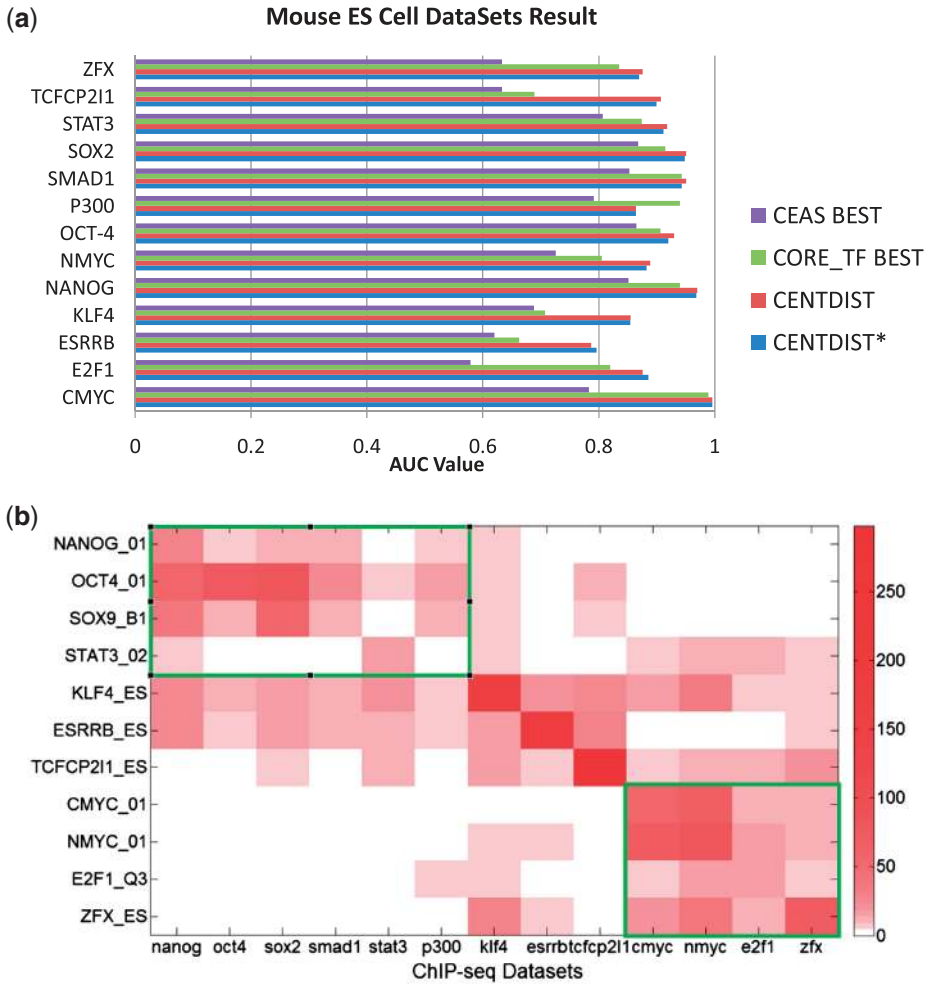

Figure 3. Co-motif analysis of 13 ES cell TFs using CENTDIST, CEAS and CORE_TF. (**a**) A comparison of co-motif analysis results using CENTDIST, CORE_TF and CEAS on 13 different ChIP-seq data sets from ES cell. The best setting in each data set for CORE_TF and CEAS were used for comparison. CENTDIST* = CENTDIST algorithm without the inclusion of velocity score. (**b**) Heat map representing the analysis of 11 ES cell core TFs motif enrichment in 13 ChIP-seq experiments. Every row corresponds to a PWM motif while every column corresponds to a ChIP-seq data set. The color of each entry presents the *Z*-score (in log scale) of the motifs with respect to the peaks of the ChIP-seq data set. The figure showed that the enhancer motifs are enriched in the enhancer ChIP-seq data sets (top-left gene rectangle) while the promoter motifs are enriched in the promoter ChIP-seq data sets (bottom-right green rectangle).

curve (AUC) (17), which ranges from 0 to 1 (a score of 0.5 is equivalent to random guessing). The details of how AUC scores were calculated can be found in Supplementary Section 2.6. Based on AUC scores, our results showed that CENTDIST significantly outperformed the best result from both CEAS and CORE_TF (Figure 3a and Supplementary Table S3). We noticed that for CEAS and CORE_TF, different configurations led to different performances, which highlights the difficulty in selecting the appropriate parameters for co-motif analysis since no single set of parameters can be considered the best for each ChIP-seq data set. CENTDIST, which requires neither background nor other parameter settings, performed significantly better (average AUC score = 0.905) than the best configuration of CEAS (average AUC score = 0.740) or CORE_TF (average AUC score = 0.84084). Furthermore, we compared the results of CENTDIST with the results ranked by frequency score only (denoted as CENTDIST* in Figure 3a and Supplementary Table S3). Overall, we found CENTDIST was better than CENTDIST* in 11 out of 13 experiments.

Next, we examined the center distribution scores of 11 ES TF motifs (Smad1 and p300 do not have known motif) across 13 TF ChIP-seq data sets (Figure 3b). From this analysis, we clearly saw two functional groups: the enhancer motifs (Oct4, Sox2, Nanog and Stat3) have good center distribution score in the enhancer TF ChIP-seq data sets (top-left green box), while the promoter motifs (cMyc, nMyc, Zfx and E2f1) have good center distribution score in the promoter TF ChIP-seq data sets (bottom-right green box) (Figure 3b). These results are in agreement with our previous findings (2). Moreover, enhancer motifs did not show good center distribution in the promoter ChIP-seq data sets, and vice versa. The only exception was Stat3, which was classified as an enhancer TF but had good center distribution at the promoter. However, a recent report showed that Stat3 was

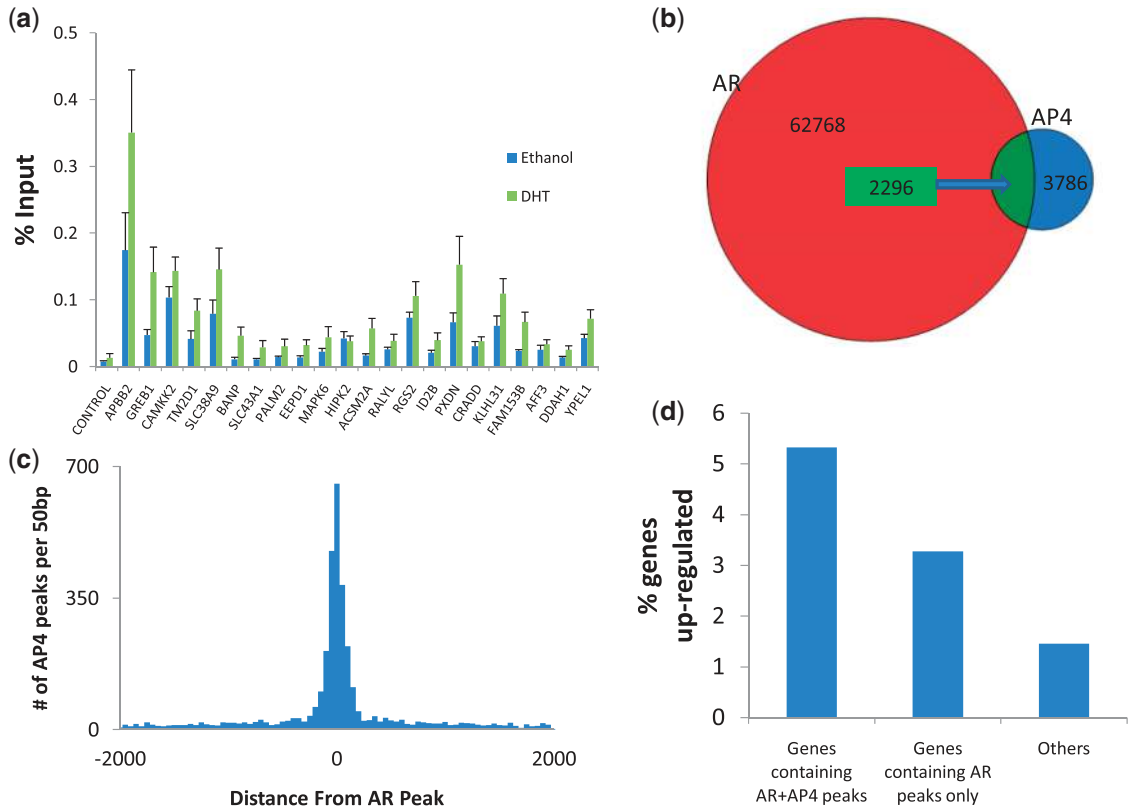

Figure 4. AP4 is a novel co-TF of AR. (**a**) ChIP-qPCR of AP4 was performed on 22 randomly selected AR peaks containing AP4 motifs in LNCaP cells before and after 2 h of DHT treatment. (**b**) Venn diagram depicting the overlap between the ChIP-seq peaks of AR and AP4. (**c**) AP4 ChIP-seq peak distribution around AR ChIP-seq peak. (**d**) Association of up-regulated genes with binding sites containing AR + AP4, AR only, or others.

also enriched in the promoter regions of ES cells, suggesting this TF can be located at both promoter and enhancer regions (18). In conclusion, the results from this large-scale comparison study demonstrate that center distribution is a good statistical model for predicting the occurrences of co-TF motifs from ChIP-seq data.

Identification of novel co-associated factors of AR in prostate cancer cells by CENTDIST

AR is a member of the nuclear hormone receptor superfamily that is important in the progression of prostate cancer (19). Recent global analyses of AR in the prostate cancer cell line, LNCaP, revealed several co-TFs (e.g. FoxA1, Oct1, Ets1, etc.) that collaborate with AR in mediating androgen-dependent transcription (1). As an independent assessment of CENTDIST performance with respect to CEAS and CORE_TF, and also to discover potential new co-TFs of AR, we compared the results of the three programs on our AR ChIP-seq data set from LNCaP cells. Again our results showed CENTDIST was the best performer among the three programs. CENTDIST identified AR and all seven known co-TFs of AR within the top 20 hits (first two columns in Supplementary Table S4). This result was significantly better than CEAS, which failed to find five of the known AR co-TFs. CORE_TF, optimized with a random background setting and 400-bp extracted-window size, identified all known AR co-TFs, however, this was within the top 37 hits. AUC analysis also indicated that CENTDIST outperformed the other motif scanners even under their best configurations (Supplementary Table S4). Furthermore, we also examine the shape of the frequency and the velocity graphs of the known co-TF motifs of AR (Supplementary Figure S3). We found that all of them have good shape even though their enrichment were not as significant as that of AR. Taken together, these results suggest that the frequency and velocity scores of co-motifs are useful information for determining the true motif signals.

Next, we validated a co-TF predicted by CENTIST. We chose a co-TF that was predicted by CENTDIST but not by the other programs. Specifically, we selected the AP4 motif, which was ranked 21st by CENTDIST. AP4 belongs to the basic helix-loop-helix (bHLH) family of TFs. It functions as a homodimer and is known to play important roles in colorectal cancer (20), however our understanding of this TF in prostate cancer is limited. To test if AP4 is a co-TF of AR, we randomly selected 22 AR ChIP-seq peaks that contain the AP4 motif and performed ChIP-qPCR in LNCaP cells treated with and without DHT. As shown in Figure 4a, all 22 binding sites showed enrichment compared to the genomic control site, suggesting that AP4 is co-localized at ARBS. Furthermore, under DHT treatment (which recruits AR), the binding of AP4 was enhanced compared to vehicle (Ethanol) treatment. To further validate whether AP4 and AR are co-binding, we took an unbiased

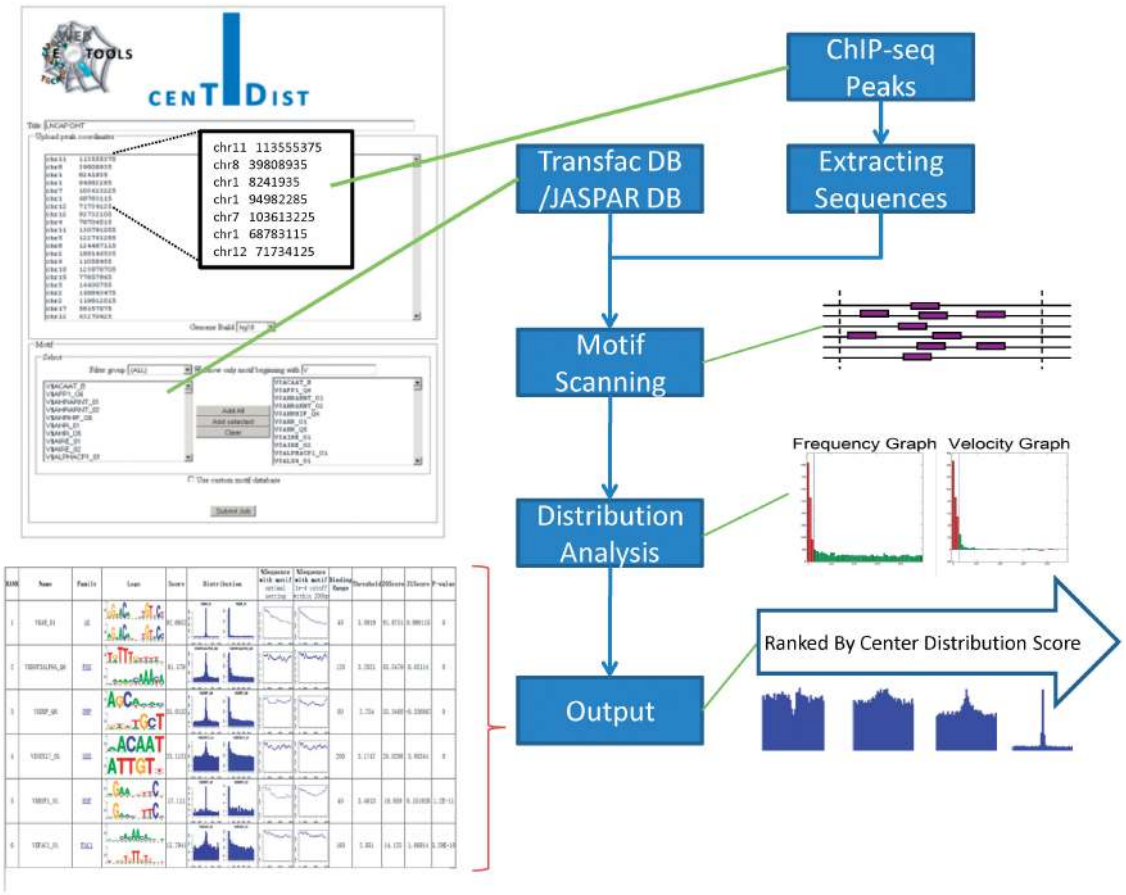

Figure 5. CENTDIST web interface and program procedure. Users can input or upload ChIP-seq peak locations or the bed format peak-region data, and select the corresponding reference genome and the motif candidates (TRANSFAC, JASPAR or custom database). After submitting the job, the data will automatically be processed according to the CENTDIST analysis pipeline. Specifically, CENTDIST will scan the sequences ($\pm$1000 bp around the peaks) and obtain the occurrences of each PWM motif to generate the frequency graph and the velocity graph. *Z*-score is used to assess the enrichment around peaks for each graph. The center distribution score of each PWM motif is calculated as the sum of the two *Z*-scores. Finally, CENTDIST outputs a list of TF families ranked by the center distribution scores.

approach and performed a ChIP-seq of AP4. As shown in Figure 4b, a large number (2296 out of 6082/38%) of AP4 ChIP-seq peaks overlapped with AR. A distribution analysis of AP4 ChIP-seq peaks around ARBS confirmed that AP4 binds in close proximity (within $\pm$200 bp) to AR (Figure 4c). We scanned for the AP4 motif in the ChIP-seq peaks and found that 79.4% of the AR-AP4 overlapping peaks contain AP4 motif. In contrast, 40.8% of the AR only peaks contain AP4 motif, however, the center distribution score for the AP4 motif around these peaks was low (Supplementary Figure S5).

Finally, we examined the fraction of androgen up-regulated genes (Supplementary Section 2.5) near AR and AP4 peaks. We divided the genes into three groups: genes with AR + AP4 peaks, genes with AR only peaks, and genes with no AR peaks. We found that the proportion of up-regulated genes in group 1 is 1.6-fold and 3.7-fold more than that in groups 2 and 3, respectively (Figure 4d), suggesting that AP4 may co-localize with AR to directly up-regulate the transcription of androgen target genes. Taken together, our results show that CENTDIST can identify novel co-TFs, even ones that are ranked very low by enrichment based methods.

CENTDIST web server implementation

Based on our observation on the imbalanced distribution of co-TF motifs around ChIP-seq peaks, we developed a web application called CENTDIST for identifying co-TFs from ChIP-seq experiments. The general pipeline of CENTDIST is shown in Figure 5.

CENTDIST is designed for the analysis high-throughput ChIP-seq data. Its web-user interface contains three main parts: input, job management and output. For input, CENTDIST accepts a list of ChIP-seq peaks. The ChIP-seq peak information can be formatted in the form of chromosome-position pairs or BED format genomic regions. CENTDIST is capable of supporting >1 million peak coordinates. The motifs used for scanning can be entered in the form of PWM or selected from either the TRANSFAC database (version 6.0), which contains 849 matrices or the JASPAR database, which has 459 matrices. CENTDIST also provides options for users to easily filter PWM motif candidates by taxonomy, or TF family. Finally, unlike other motif scanning programs, CENTDIST is totally parameter free. Users are not required to provide the

background, the enrichment window size or even choose the FDR or PWM cut-off for the PWM motifs. All these parameters will be estimated by CENTDIST automatically.

With regards to job management, submitted jobs will be sent to the job queue on the server and processed based on a first come first serve policy. Users can view the status of their submitted jobs, and access or delete the results of previous runs at the 'Job Management' page (Supplementary Section 5.4). The page refreshes automatically and email notifications will be sent to users once the jobs are completed.

The main output page for CENTDIST is a table containing PWM motifs ranked according to center distribution scores (Supplementary Section 5.5). Each row in the table presents the enriched TF family, and user can click on a link associated with each TF family to browse the result of each individual member. The output also contains visualization features like the PWM logo (21) of the motif, the frequency graph (center view and folding view), and other useful numeric features like binding range (the enrichment window size), PWM score threshold (the cut-off that maximizes the center distribution score), center distribution score and *P*-value. In addition, the output page provides the motif distribution across different peak ranks (columns 7 and 8 in Supplementary Figure S12), which is useful when the input peaks are sorted by some quality measure like ChIP-seq intensity.

DISCUSSION

The performance of existing motif scanners is heavily dependent on selecting the proper background and other parameter settings. Choosing the correct background, however, is currently considered an art. What is more, there is no one set of parameters that can satisfy all co-TFs. Finally, the assumption that noise is uniformly distributed may not be true when CG (or AT) content varies in ChIPed enriched regions. In this article, we present a new computational method called CENTDIST that utilizes frequency information as well as slope information (velocity) to predict whether a motif is real or not. CENTDIST does not require an explicit background model. Using the velocity score, CENTDIST is also insensitive to CG- or AT-biases. We examined CENTDIST on 14 ChIP-seq data sets and demonstrated that it is better than existing methods. We also show that this can be achieved without requiring expert knowledge in configuring the program. For the ChIP-seq of in LNCaP cells, CENTDIST discovered AP4 as a novel co-TF of AR, which was missed by existing enrichment-based methods. CENTDIST also discovered nine additional co-TF motifs that were unique to the program. For five of these co-TF motifs, evidence from literature suggests that they could be the motifs of potential collaborators of AR (Supplementary Table S5).

CENTDIST does have certain limitations. For example, CENTDIST may fail to identify co-TFs whose binding site distribution does not follow the proximity assumption (i.e. co-TFs that are not co-localized with the ChIPed TF). However, the latter would not be found by traditional enrichment-based methods either since their binding sites are not enriched.

CENTDIST is a user-friendly web-based application that is capable of analyzing large-scale ChIP-seq data sets. It can scan ~700 TRANSFAC motifs over a ChIP-seq data set containing 10 000 peaks in only 10 min (Supplementary Table S9). With CENTDIST, users do not have to set any parameters except to upload the ChIP-seq peak locations and select the PWM motif library they wish to use for scanning. The output of CENTDIST contains clean and rich information for users. Specifically, it groups the list of enriched motifs into TF families, and provides other information including PWM logo, motif distribution graph, enrichment *P*-value and the enriched window size of the enriched motifs.

To the best of our knowledge, CENTDIST is the first motif scanner for ChIP-seq data that utilizes the shape information (velocity) of the motif distribution, and automatically detects the size of the motif-enriched region and PWM score cut-off. It compares the enrichment inside/outside of the enriched motif region without the input of additional background information. Although there is still room to improve the methodology, this study opens a new door for utilizing the shape information to extract biologically meaningful co-TFs in a ChIP-seq data set.

ACCESSION NUMBER

GSE28264, GSE28857, GSE28596.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: Biomedical Research Council/Science and Engineering Research Council of A*STAR (Agency for Science, Technology and Research), Singapore (in part); MOEs AcRF Tier 2 funding R-252-000-444-112; A*STAR graduate scholarship (to C.W.C., in part); National University of Singapore research scholarship (to Z.Z.Z., in part).

Conflict of interest statement. None declared.

REFERENCES

1. Cheung,E. and Kraus,W.L. (2010) Genomic analyses of hormone signaling and gene regulation. *Ann. Rev. Physiol.*, **72**, 191–218.
2. 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.
3. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, **316**, 1497–1502.
4. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.*

(2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
5. Xu,H., Handoko,L., Wei,X., Ye,C., Sheng,J., Wei,C.L., Lin,F. and Sung,W.K. (2010) A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics (Oxford, England)*, **26**, 1199–1204.
6. Ji,X., Li,W., Song,J., Wei,L. and Liu,X.S. (2006) CEAS: cis-regulatory element annotation system. *Nucleic Acids Res.*, **34**, W551–W554.
7. Hestand,M.S., van Galen,M., Villerius,M.P., van Ommen,G.J., den Dunnen,J.T. and t Hoen,P.A. (2008) CORE_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated genes. *BMC Bioinformatics*, **9**, 495.
8. Hooghe,B., Hulpiau,P., van Roy,F. and De Bleser,P. (2008) ConTra: a promoter alignment analysis tool for identification of transcription factor binding sites across species. *Nucleic Acids Res.*, **36**, W128–W132.
9. Ho Sui,S.J., Fulton,D.L., Arenillas,D.J., Kwon,A.T. and Wasserman,W.W. (2007) oPOSSUM: integrated tools for analysis of regulatory motif over-representation. *Nucleic Acids Res.*, **35**, W245–W252.
10. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, **16**, 16–23.
11. He,H.H., Meyer,C.A., Shin,H., Bailey,S.T., Wei,G., Wang,Q., Zhang,Y., Xu,K., Ni,M., Lupien,M. *et al.* (2010) Nucleosome dynamics define transcriptional enhancers. *Nature Genet.*, **42**, 343–347.
12. Wederell,E.D., Bilenky,M., Cullum,R., Thiessen,N., Dagpinar,M., Delaney,A., Varhol,R., Zhao,Y., Zeng,T., Bernier,B. *et al.* (2008) Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.*, **36**, 4549–4564.
13. Sharov,A.A. and Ko,M.S. (2009) Exhaustive search for over-represented DNA sequence motifs with CisFinder. *DNA Res. Int. J Rapid Public. Reports Gene Genome*, **16**, 261–273.
14. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
15. Sandelin,A., Alkema,W., Engstrom,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.
16. Raha,D., Wang,Z., Moqtaderi,Z., Wu,L., Zhong,G., Gerstein,M., Struhl,K. and Snyder,M. (2010) Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc. Natl Acad. Sci. USA*, **107**, 3639–3644.
17. Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.
18. Kidder,B.L., Yang,J. and Palmer,S. (2008) Stat3 and c-Myc genome-wide promoter occupancy in embryonic stem cells. *PloS One*, **3**, e3932.
19. Heinlein,C.A. and Chang,C. (2004) Androgen receptor in prostate cancer. *Endocr. Rev.*, **25**, 276–308.
20. Cao,J., Tang,M., Li,W.L., Xie,J., Du,H., Tang,W.B., Wang,H., Chen,X.W., Xiao,H. and Li,Y. (2009) Upregulation of activator protein-4 in human colorectal cancer with metastasis. *Int. J. Surg. Pathol.*, **17**, 16–21.
21. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.