doi:10.20944/preprints202007.0639.v1

Article

Asymmetric Conservation within Pairs of Co-occurred Motifs Mediates Weak Direct Transcription Factor Binding in ChIP-seq Data

Victor Levitsky [1,2]*, Dmitry Oshchepkov [1] , Elena Zemlyanskaya [1,2] and Tatyana Merkulova [1,2]*

[1] Department of System Biology, Institute of Cytology and Genetics, Novosibirsk, 630090, Russia
[2] Department of Natural Science, Novosibirsk State University, Novosibirsk, 630090, Russia
* Correspondence: levitsky@bionet.nsc.ru

Abstract: (1) Background: Transcription factors (TFs) are main regulators of eukaryotic gene expression. The cooperative binding to genomic DNA of at least two TFs is the widespread mechanism of transcription regulation. Cooperating TFs can be revealed through the analysis of co-occurrence of their motifs. (2) Methods: We applied Motifs Co-Occurrence Tool (MCOT) that predicted pairs of spaced or overlapped motifs (composite elements, CEs) for a single ChIP-seq dataset. We improved MCOT capability for prediction of asymmetric CEs with one of participating motifs possessing higher conservation than another does. (3) Results: Analysis of 119 ChIP-seq datasets for 45 human TFs revealed that almost for all families of TFs the co-occurrence with an overlap between motifs of target TFs and more conserved partner motifs was significantly higher than that for less conserved partner motifs. The asymmetry toward partner TFs was the most clear for partner motifs of TFs from ETS family. (4) Conclusion: Co-occurrence with an overlap of less conserved motif of a target TF and more conserved motifs of partner TFs explained a substantial portion of ChIP-seq data lacking conserved motifs of target TFs. Among other TF families, conservative motifs of TFs from ETS family were the most prone to mediate interaction of target TFs with its weak motifs in ChIP-seq.

Keywords: Chromatin immunoprecipitation followed by sequencing, transcription factors binding sites prediction, cooperative binding of transcription factors, composite elements, motifs conservation, classification of transcription factors, ETS transcription factor family, direct binding of transcription factors, overlap of motifs.

1. Introduction

Tissue-, cell- and stage-specific regulation of gene expression is produced through interactions of transcription factors (TFs) with respective regulatory elements called binding sites (BSs) or motifs. Typically, each TF functions in tight cooperation with other TFs: there is a variety of mechanisms for cooperative TF-DNA binding [1,2]. Roughly, these mechanisms may be classified into simultaneous and sequential. The first option implies a protein-protein interaction, and subsequent homo- or heterodimer binding to DNA. This mechanism may allow comparable or approximately equal impacts of affinity of two respective motifs. Alternatively, one TF of a pair may preliminary interact with DNA, and at the second stage ternary complex is formed through contributions of protein-protein and protein-DNA contacts of the second TF. This opportunity is facilitated by a higher DNA affinity of the first TF than for the second one. DNA-mediated interaction may also be facilitated by DNA conformation or nucleosomal organization. E.g., the propensity to interact with nucleosomal DNA is a special mark of pioneer TF [3-5]. Thus, various mechanisms may explain a variety of possible TF-DNA ternary complexes, but in many cases, we may expect that behavior of two TFs is asymmetric. The recent review [6] proposed that in co-occurred pairs of motifs besides the orientation and spacing, the strength (affinity) of the individual motifs contributes to the specificity of a DNA regulatory region. Hence, systematic analysis of all possible partner motifs for various target motifs may propose the possible mechanism of cooperative TFs action.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) analysis became the gold standard for protein/DNA-binding annotation at the whole genome level. In particular, ChIP-seq approach has been widely applied for annotation of TFBSs. The standard analysis pipeline at the final stage proposed application of *de novo* motif search tool that could confirm the presence of BSs (motifs) specific for target (anchor) TF [7]. Since application of *de novo* motif search tools for a single ChIP-seq datasets become a routine procedure [8], several attempts underlined the importance of massive analysis of motifs co-occurrence that reflected the cooperative mechanisms of TF actions [9,10]. We recently proposed the Motifs Co-Occurrence Tool (MCOT) package for motifs co-occurrence prediction in ChIP-seq data [11]. MCOT possessed two specific features which were absent in other analogous bioinformatics tools. First, MCOT used a single ChIP-seq dataset for discovering motifs co-occurrence with a spacer and with an overlap. Second, MCOT performed simultaneous application of several thresholds for each motif; consequently, MCOT was able to retrieve CEs of anchor and partner motifs with various conservation ratios. Here the conservation of a motif implies its similarity to a recognition model.

In the current study we aimed with a benchmark ChIP-seq data for various TFs to map respective anchor motifs and estimate which potential partner TFs through co-occurrence of their motifs with motifs of anchor TFs might mediate binding of anchor TFs. In particular, we asked whether asymmetric pairs of anchor and partner motifs with more conserved partner motifs could explain earlier known substantial portions of ChIP-seq data lacking conserved anchor motifs (about a half of a ChIP-seq dataset, [12]. To investigate this issue we proposed the improvement for the MCOT computation procedure that directly reflected whether an observed misbalance between conservation of anchor and partner motifs was significantly higher than a random expectation. Consequently, for each CE consisting of anchor and partner motifs, beside the conventional significance that reflected it significance, MCOT provided two additional significances that reflected the enrichments of asymmetric CEs with more conserved anchor and partner motifs.

We carefully annotated anchor motifs for benchmark ChIP-seq data. Next, we calculated the abundance of CEs with a spacer and with an overlap of anchor and partner motifs from a library of previously known potential partner motifs. In particular, for each partner motif we separately analyzed asymmetric CEs with higher and lower conservation of partner motifs than respective anchor motifs. Finally, we classified all partner motifs according to families of partner TFs [13].

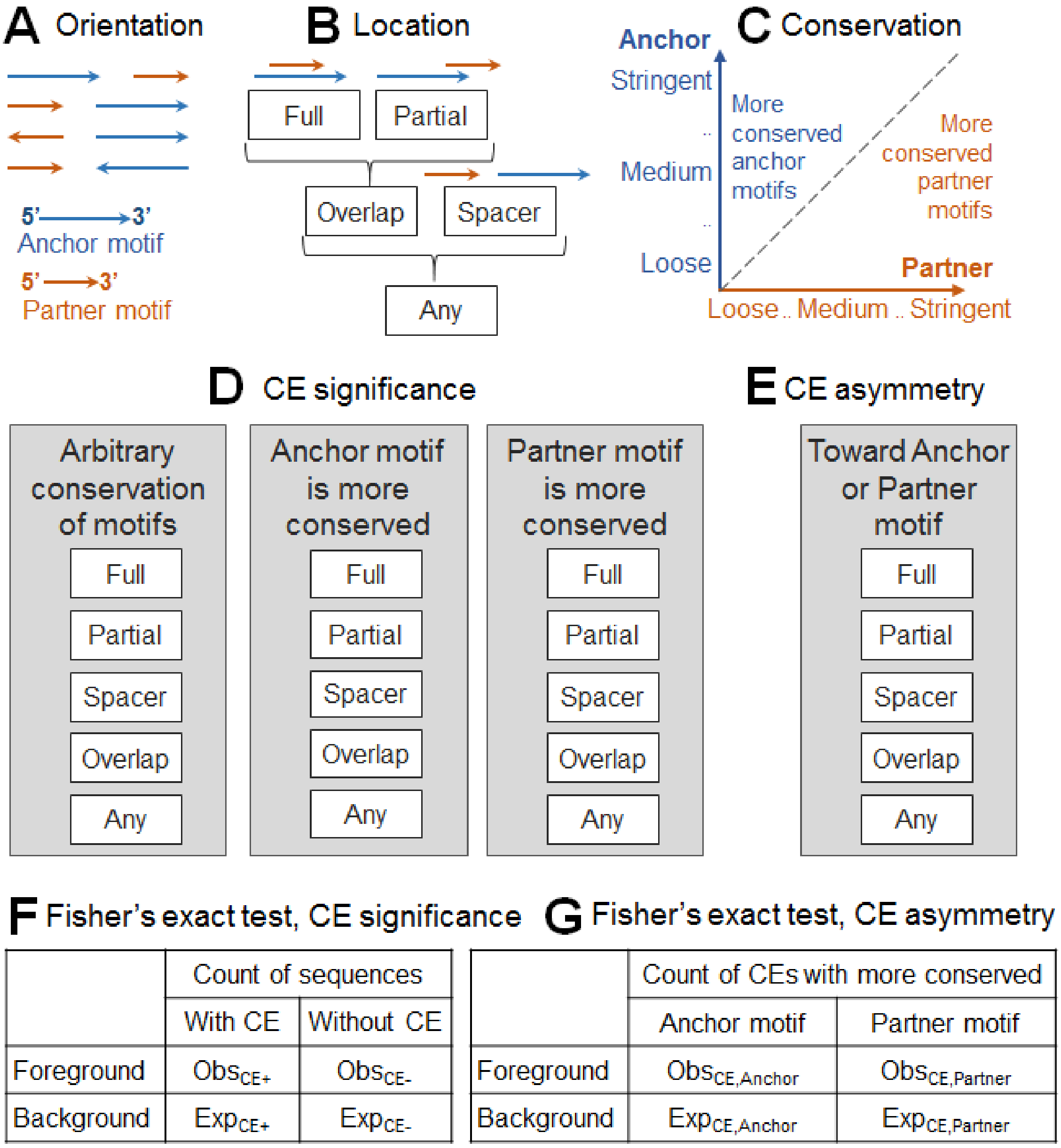
We concluded that only among overlapping pairs of anchor and partner motifs respecting all families of partner TFs, pairs with higher conservation of partner motifs were significantly more abundant than those with higher conservation of anchor motifs. Various TF families were differentiated according to the misbalance between asymmetric CEs with more conserved anchor and partner motifs. Thus, overrepresented asymmetric CEs with more conserved partner motifs and less conserved anchor motifs systematically promoted weak direct interactions of anchor TFs in ChIP-seq data. Among other families, partner motifs of TFs from ETS family had the greatest misbalance in conservation toward partner motifs. Hence, we have shown that motifs of TFs from ETS family systematically mediate cooperative binding of other TFs through higher conservation of Ets-like motifs in widespread CEs with an overlap of motifs.

2. Results

2.1. Integration of CE significance and CE asymmetry in MCOT analysis

doi:10.20944/preprints202007.0639.v1

We earlier developed MCOT package for prediction of spaced and overlapped pairs of co-occurred motifs for a single ChIP-seq dataset [11]. To perform the search of CEs, MCOT required the ChIP-seq dataset (peaks), the anchor motif that refers to target TF, and either the partner motif or the assignment of a public library of proven partner motifs, e.g. Hocomoco [14]. In the current study, we applied the model of Position Weight Matrix (PWM) for motifs recognition. Besides the classification of CEs by the orientation, MCOT classified them into fully/partially overlapped and spaced (Figure 1A,B).


A Orientation
5'→3'
Anchor motif
5'→3'
Partner motif
B Location
Full
Partial
Overlap
Spacer
Any
C Conservation
Anchor
Stringent
Medium
Loose
More conserved anchor motifs
More conserved partner motifs
Partner
Loose .. Medium .. Stringent
D CE significance
Arbitrary conservation of motifs
Anchor motif is more conserved
Partner motif is more conserved
Full
Partial
Spacer
Overlap
Any
E CE asymmetry
Toward Anchor or Partner motif
Full
Partial
Spacer
Overlap
Any
F Fisher's exact test, CE significance
G Fisher's exact test, CE asymmetry

	Count of sequences	
	With CE	Without CE
Foreground	Obs_{CE+}	Obs_{CE-}
Background	Exp_{CE+}	Exp_{CE-}

	Count of CEs with more conserved	
	Anchor motif	Partner motif
Foreground	$Obs_{CE,Anchor}$	$Obs_{CE,Partner}$
Background	$Exp_{CE,Anchor}$	$Exp_{CE,Partner}$

Figure 1. MCOT package basic concept and output data. MCOT classified all CEs according to mutual orientation (panel A), location (B) and conservation of two motifs (C). The next analysis was subdivided on five computation flows (Full, Partial, Overlap, Spacer and Any), that were defined according mutual location of motifs (B). The significance of CEs was computed without respect to relative conservation of motifs in a pair, or either an anchor or a partner motif should have more conserved hits (D). The asymmetry for a certain CE could be significant either toward anchor or partner motif. MCOT constructed 2x2 contingency tables for calculation of significance of CE (F) and CE asymmetry (G). To compute the significance of CEs (F) peaks with hits of both motifs were left in analysis and MCOT compared fractions of peaks/permuted sequences that contained either any CEs, or only CEs with more conserved anchor or partner motifs (C). To compute the significance of

asymmetry within CE (G) MCOT compiled the list of all CEs in peaks and permuted sequences and compared respective amounts of CEs with more conserved anchor motif and the rest CEs with more conserved partner motif.

In the current study, we updated the CEs classification according to motifs conservation. The analysis of the scatterplot between conservation of anchor and partner motifs may reveal an extent of misbalance between similarity to recognition models of their motifs (Figure 1C), more specifically the value -Log$_{10}$(FPR) is the measure of motif's conservation, here FPR denotes False Positive Rate, see Materials and Methods).

Basic MCOT output data represented the significance of CEs without respect to conservation of motifs in a pair and those for CEs with more conserved anchor or partner motifs (Figure 1D). Additionally, for each pair of motifs MCOT computed the significance of CE asymmetry (Figure 1E). Panels F and G show 2x2 contingency tables that illustrate the application of Fisher's exact test. In particular, both types of test compared the foreground and background data. The test of CE significance (Figure 1F) implied the comparison of portions of sequences with CEs among all sequences with hits of both participating motifs. The test of CE asymmetry (Figure 1G) compared the fractions of CEs that had certain one motif more conserved than another did. In particular, we performed a triple application of similar Fisher's exact tests from Figure 1F. The first test estimated peaks with CEs possessing arbitrary relationships between motifs conservation (the area of the whole square in Figure 1C). The second and third tests computed the significance for asymmetric CEs, i.e. we took in account only peaks containing CEs from the upper left and lower right triangles (Figure 1C), that respected to asymmetric CEs toward the anchor and asymmetric CEs toward partner motifs.

2.2. Single ChIP-seq dataset: example of significant asymmetry within CE

In this section we illustrate calculation of CE asymmetry (see Figure 1E,G and Materials and Methods) with ChIP-seq dataset of FoxA2 from mouse liver tissue [15]. In this study, besides the enrichment of anchor FoxA2 motifs, authors revealed its co-occurrence with motifs of partner TFs GATA4, PAX6 and HNF1. MCOT analysis confirmed the co-occurrence of motifs of these TFs with FoxA2 motifs (respective partner motifs were taken from the Hocomoco mouse core collection, [14]. But, only for HNF1β motif (HNF1B_MOUSE.H11MO.0.A) we found the extremely significant asymmetry within predicted CEs toward partner motif ($p < 2E-28$ and $p < 4E-17$ for CEs with an overlap of motifs and with a spacer, respectively). Figure 2 shows the difference between relative frequencies of observed and expected CEs with specific conservation of FoxA2 and HNF1β motifs for their overlapped and spaced positioning. MCOT analysis of other ChIP-seq datasets for FoxA2 and it close homologue FoxA1 revealed that FoxA1/2-HNF1β CEs with an overlap of motifs were always significant, in some cases a moderate significance also found for respective CEs with a spacer, but the significant asymmetry in these CEs was not observed for all other FoxA1/2 ChIP-seq datasets (FoxA2 for liver cell line HepG2 [16]; GSM686926, FoxA1, prostate cell line LNCaP [17]; GSM1505633, FoxA1, embryonic cell lines, [18]). Thus, CE asymmetry toward HNF1β motif appeared to be the specific feature of FoxA2-HNF1β CEs in liver tissue.

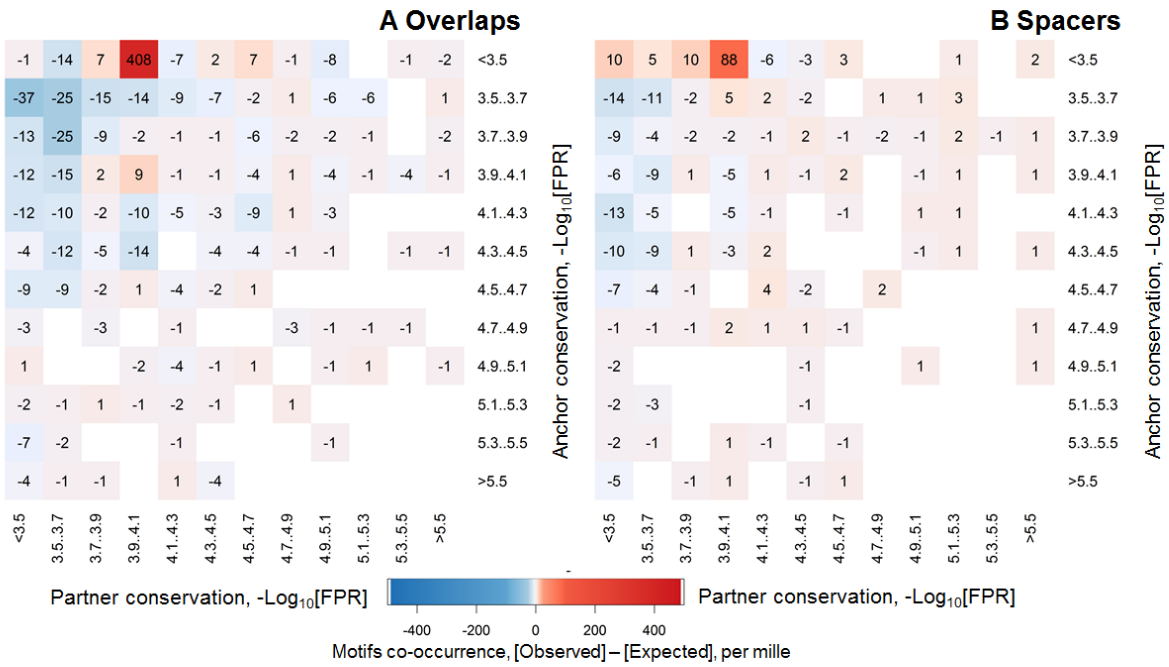


Figure 2. The difference between observed and expected abundances of CEs with specific conservation of the anchor FoxA2 (axis Y) and partner HNF1β (axis X) motifs for ChIP-seq data from liver tissue [15], in per mille. The conservation of motifs was measured as an expected occurrence, -Log$_{10}$(FPR) (see Materials and Methods). The color on both heatmaps shows the difference between observed (peaks) and expected (permuted sequences) relative abundance of CEs with specific conservation levels (see Materials and Methods). FoxA2 and HNF1β motifs were derived from Homer de novo motif search [8] and Hocomoco database (HNF1B_Mouse.H11MO.0.A, [14]. Panels (A) and (B) show asymmetry of CEs with an overlap of motifs and with a spacer, respectively.

2.3. Single ChIP-seq dataset: multiple partner TFs support binding of anchor TF

The application of MCOT may provide a list of potential partner motifs with the designation of relationship between conservation of motifs in a pair. The previous section represented a sole example of partner motif that had higher conservation than the motif of anchor TF. Since the database of well-annotated potential partner motifs was recently developed [14], in practice, multiple significant asymmetric CEs of various partner TFs may cooperate with an anchor TF, and this may be a possible explanation of the discrepancy between the total number of peaks and the presence of respective BSs of an anchor motif in these peaks [12,19]. We previously showed that at least for FoxA2 in concrete ChIP-seq datasets [15,16] almost 100% of peaks contained potential motifs of anchor TF, although this conclusion was deduced due to alternative to PWM recognition model [20]. Consequently, the majority of FoxA2 peaks should contain at least moderately conserved FoxA2 motifs. Hence, we excluded from analysis 37.71% of all (4455) FoxA2 peaks that had the most conservative hits (FPR < 5.24E-5) and 19.08% of peaks that had too weak conservation of FoxA2 best hits (FPR > 5E-4), see Materials and Methods. The rest 43.21% of peaks had FoxA2 hits with a moderate or weak conservation, 5.24E-5 < FPR < 5E-4. Probably, a portion of these peaks contained CEs respecting to various more conserved partner motifs. Consequently, we required that each partner motif beside the absence of similarity to an anchor motif should have the significance of asymmetric CEs toward partner motif. We selected the top 30 such motifs according to respective CE significance and sorted them according to the fraction of peaks containing asymmetric CEs with more conserved partner motifs (see Figure 1D,F and Materials and Methods). Figure 3A represents the ranging of these partner motifs. Almost all peaks of the fraction with moderately and weakly conserved FoxA2 hits (89.6%) contained significant CEs with more conserved partner motifs. The first ranked motif FoxQ1 belonged the same family as the FoxA2 motif (Forkhead box (FOX) factors{3.3.1}, these motifs were

doi:10.20944/preprints202007.0639.v1

moderately similar (p<0.1), i.e. the filter detected their homology as not significant. Among the top-ranked partner motifs we found the BSs for previously known co-factors HNF1α/β and HNF6 (Figure 3A). The similarity filter excluded from our analysis motif HNF4γ (HNF4G_MOUSE.H11MO.0.C). Analysis of motifs similarity within top 30 partner motifs (Figure 3B) demonstrated that these motifs respected to relatively small numbers of TF families [13]. Thus, besides the first ranked FoxQ1 that belonged the family Forkhead box (FOX) factors{3.3.1}, next eight top-ranked motifs belong to five families:

- Thyroid hormone receptor-related factors (NR1){2.1.2} (NR1H3_MOUSE.H11MO.0.A),
- POU domain factors{3.1.10} (HNF1A_MOUSE.H11MO.0.A and HNF1B_MOUSE.H11MO.0.A),
- HD-CUT factors{3.1.9} (HNF6_MOUSE.H11MO.0.A and CUX2_MOUSE.H11MO.0.C),
- C/EBP-related{1.1.8} (NFIL3_MOUSE.H11MO.0.C),
- SOX-related factors{4.1.1} (SOX9_MOUSE.H11MO.0.A and SOX10_MOUSE.H11MO.0.B).

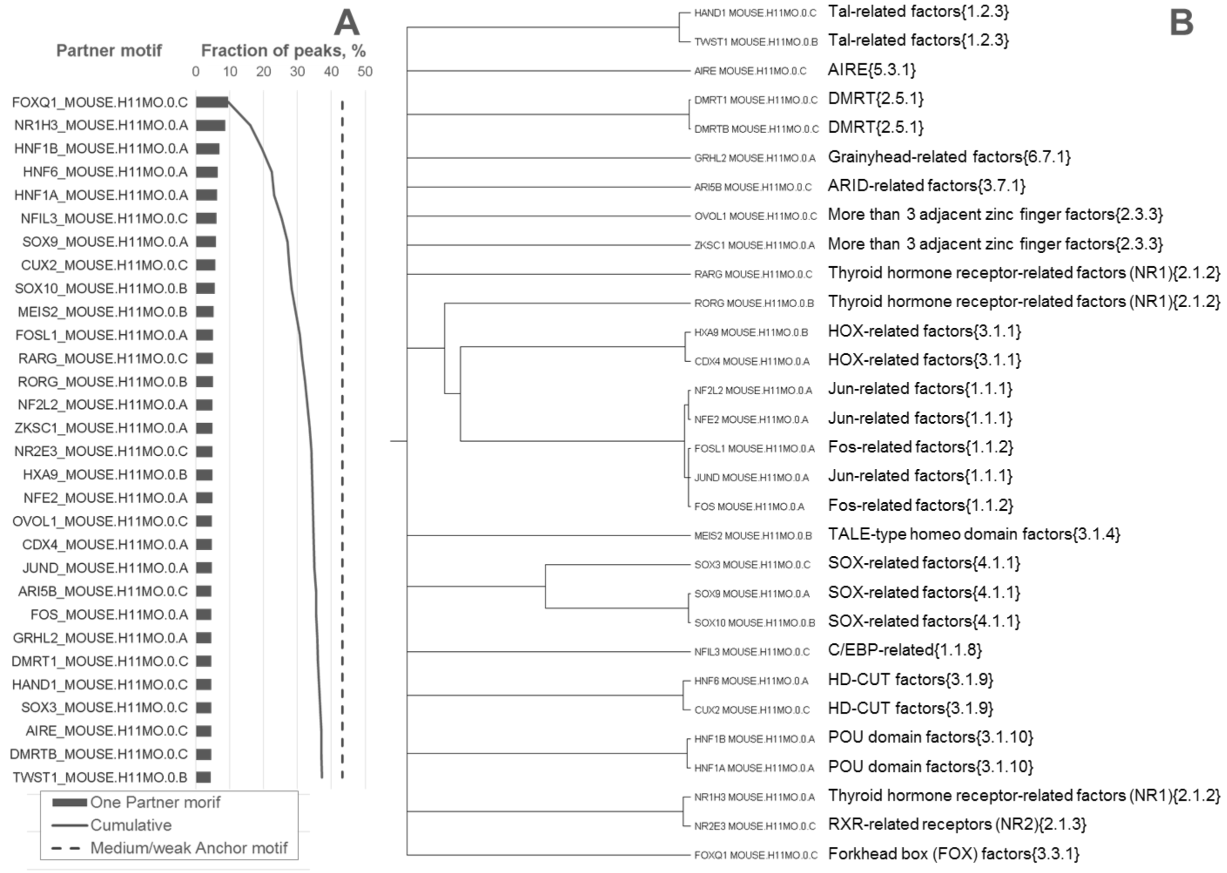


Figure 3. The analysis of FoxA2 peaks [15] that contained the FoxA2 motifs of moderate and weak conservation. Panel (A) displays fractions of analyzed peaks that contained asymmetric CEs with specific partner motifs. Panel (B) shows the tree of similarity for selected list of 30 top-scored partner motifs from panel A and the respective families of partner TFs [13]. We took in analysis only 43.21% of all peaks with best scores of peaks in the range of FPR from 5.24E-5 to 5E-4, see the dashed line in panel (A). We applied MCOT package and defined top-ranked 30 partner motifs from Hocomoco mouse core collection [14] that did not have similarity to the anchor motifs (p < 0.05, any similarity measure from [11] and respected the significant asymmetric CEs toward the partner motif. We sorted partner motifs according to the fraction of peaks that contained such asymmetric CEs (panel (A), columns) and computed the cumulative fraction of peaks that contained at least one, two, three, etc. up to 30 types CE types with various top-scored partner motifs, see the regular line in panel (B).

Notably, in the more common variants of asymmetric CEs FoxA2/HNF1β, FoxA2/HNF6 and FoxA2/Sox9 were used very different structural types of FoxA2, which could be represented by consensuses TATTTATTTA, TATTGACT and TGTTT(A/G)(C/T) (Figure S1), i.e. in each case the motif of target TF is adopted by the respective partner motif.

In total, these ten top-ranked asymmetric CEs with more conserved partner motifs contained in 29.5% of all ChIP-seq peaks and in 68.2% of peaks with moderately or weakly conserved FoxA2 motifs with FPRs from 5.24E-5 to 5E-4 (Figure 3). 30 top-ranked asymmetric CEs increased these fractions up to 37.3% and 86.4%, respectively. Thus, a substantial portion of the ChIP-seq dataset contained asymmetric CEs with more conserved partner motifs.

2.4. Massive analysis of asymmetric CEs

2.4.1. Analysis of partner motifs classified according to TFs families

In the previous section we performed analysis of a single ChIP-seq dataset and approved that multiple motifs of various known and presumed partner TFs might be located near weak motifs of anchor TF (Figure 3). Since we had benchmark ChIP-seq data [11] supplied with results of *de novo* motif search [8] for various TFs in different tissues, development stages and conditions, we asked whether certain partner motifs tended to mediate binding of many various anchor TFs through asymmetric CEs with more conserved motifs of partner TFs. We took in analysis the benchmark data of 119 ChIP-seq datasets for 45 human TFs with annotated occurrences of anchor motifs and applied the library of 396 partner motifs from Hocomoco database (see Materials and Methods). We applied MCOT package for prediction of CEs without reference to relationships of anchor and partner motifs conservation, and of CEs either with more conservative anchor or partner motifs (see Materials and Methods and Figure 1D). Abundances of these types of CEs for Full, Partial, Overlap, Spacer and Any computation flows for all partner motifs are given in Table S2.

Since MCOT operated motifs, but not TFs, and the Hocomoco collections contained hundreds of more or less homologous motifs of various TFs, but earlier all human TFs were classified by their DNA-binding domains [13], we classified all accepted in analysis 396 partner motifs into 50 clades. These clades comprised 49 families of TFs, and also one additional subfamily of CTCF-like motifs according to previous results [12] (see Materials and Methods). Thus, we asked whether more conserved partner motifs from specific families of TFs could systematically form CEs with anchor motifs from various ChIP-seq datasets.

The total counts of asymmetric CEs for all anchor and partner motifs were 4228 and 14718 for Full and Overlap computation flows, respectively; the rest flows revealed substantially lower amounts (519, 56 and 15 for Partial, Spacer and Any flows, respectively; Table S2). The Welch's t-test for the distribution of all partner motifs between the number of ChIP-seq datasets for asymmetric CEs toward the partner motifs vs. that for asymmetric CEs toward anchor motifs demonstrated the significance for Full, Partial and Overlap flows ($p < 0.02$, $p < 1E-34$ and $p < 1E-170$, respectively; Table S2). The computation flows Spacer and Any revealed the significance in reverse direction, i.e. asymmetric CEs toward the anchor motifs were more abundant than those toward the partner motifs ($p < 1E-9$ and $p < 1E-41$, respectively; Table S2). Thus, the higher conservation of partner motifs in CEs with an overlap of motifs have systematic behavior and abundance of such asymmetric CEs was substantially higher than that for others flows. Hence, the focus in the consequent analysis will be on CEs with an overlap of motifs.

Figure 4 compares the number of ChIP-seq datasets containing asymmetric CEs toward partner motifs and overlaps of motifs and respective number for asymmetric CEs toward anchor motifs for 50 selected above clades of TFs for the benchmark ChIP-seq data.

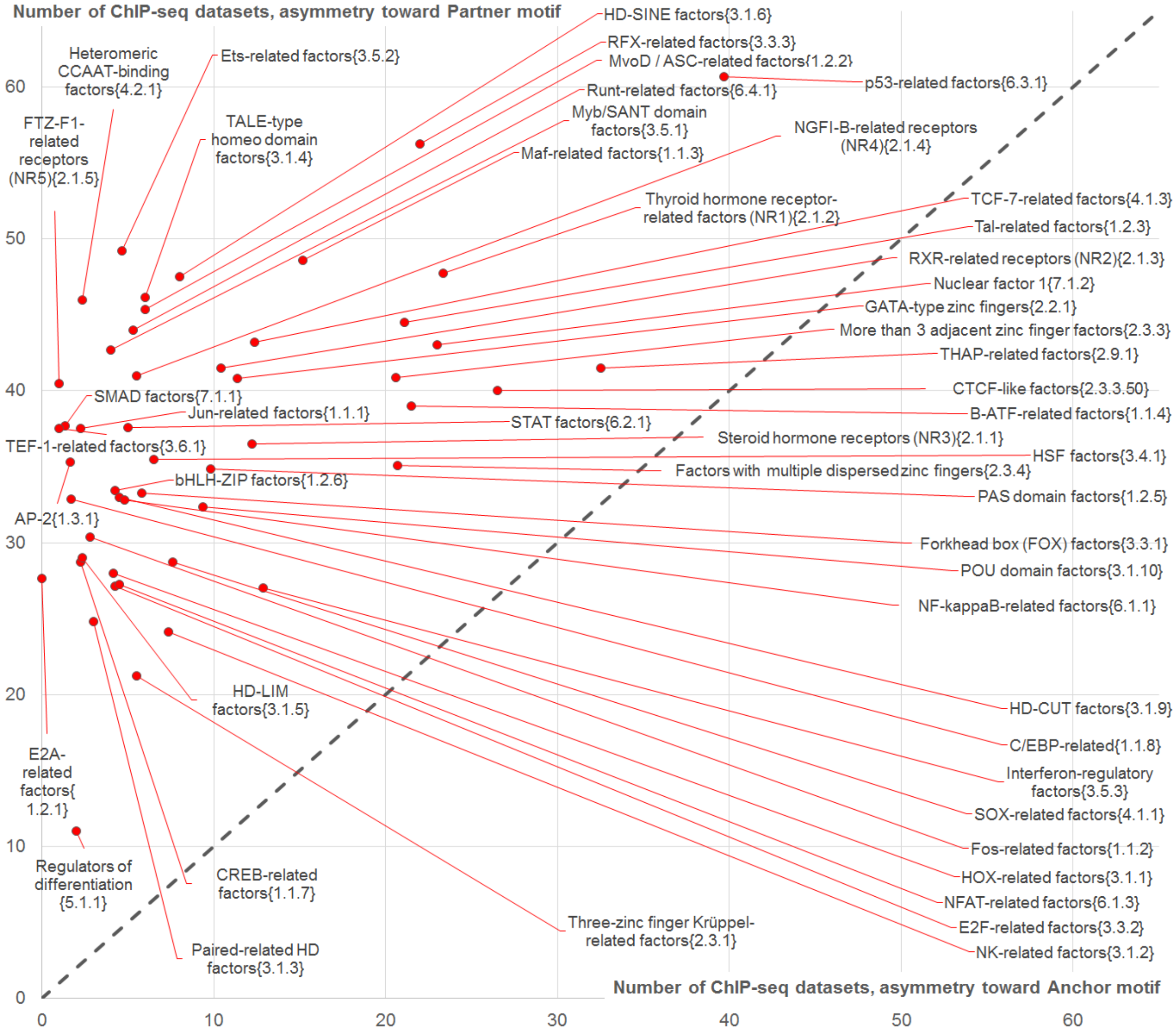


Figure 4. The scatterplot of abundances of asymmetric CEs toward the anchor (axis X) and asymmetric CEs toward the partner (axis Y) motifs for 50 clades of TFs. These clades comprised 49 families of TFs with at least two motifs and subfamily CTCF-like factors{2.3.3.50} with two motifs from the Hocomoco human core collection [14]. Total number of ChIP-seq datasets was equal to 119 (see Materials and Methods). All predicted CEs respected the computation flow Overlap. The diagonal dashed line marks equal numbers of datasets, it implies the partitioning of all clades into those with the higher abundance of asymmetric CEs toward partner motifs (top left triangle, all 50 clades) and those with the higher abundance of asymmetric CEs toward anchor motifs (bottom right triangle without clades).

We concluded that for all clades of partner motifs the abundance of CEs with asymmetry toward partner motifs exceeded that with asymmetry toward anchor motifs, since for all clades of TFs respective points in Figure 4 lie above the diagonal from the lower left to upper right (dashed line). The most specific clades of partner TFs for asymmetry toward partner motifs respected to families of Ets-related factors{3.5.2} and Heteromeric CCAAT-binding factors{4.2.1} (see points close to the top left corner, Figure 4). While families of p53-related factors{6.3.1}, RFX-related factors{3.3.3}, and Thyroid hormone receptor-related factors, (NR1){2.1.2} showed high abundance of both asymmetric CEs toward the anchor and partner motifs, since their points were close to the diagonal in Figure 4. THAP11 and CTCF-like motifs, among motifs for TFs from other clades, had tendency to form asymmetric CEs toward the anchor motifs, since top six clades for asymmetric CEs toward the anchor motifs were p53-related factors{6.3.1}, THAP-related factors{2.9.1}, CTCF-like factors{2.3.3.50}, Thyroid hormone receptor-related factors (NR1){2.1.2}, Nuclear factor 1{7.1.2} and RFX-related factors{3.3.3} (Figure 4).

To estimate for motifs overlapping the enrichment of asymmetric CEs toward partner motifs vs. those asymmetric toward anchor motifs we applied the Welch's t-test for the number of ChIP-seq datasets. Figure 5 shows the dependence of the significance of t-test for the number of ChIP-seq datasets containing CEs with an overlap of motifs (and without taking into account motifs conservation). Application of Bonferroni's correction threshold for the significance of difference, p < 0.05/50 = 0.001 (Figure 5) have shown that 45 out of all 50 clades (90%) possessed the significant increase of the number of ChIP-seq datasets with asymmetric CEs toward the partner motifs vs. that with asymmetric CEs toward the anchor motifs. We found that the Ets-related factors{3.5.2} family combined

- high abundance of CEs (axis X in Figure 5),
- significant enrichment of asymmetric CEs toward partner motifs vs. those asymmetric toward anchor motifs (axis Y in Figure 5);
- high abundance of asymmetric CEs toward partner motifs in comparison with that for asymmetric CEs toward anchor motifs (Figure 4).

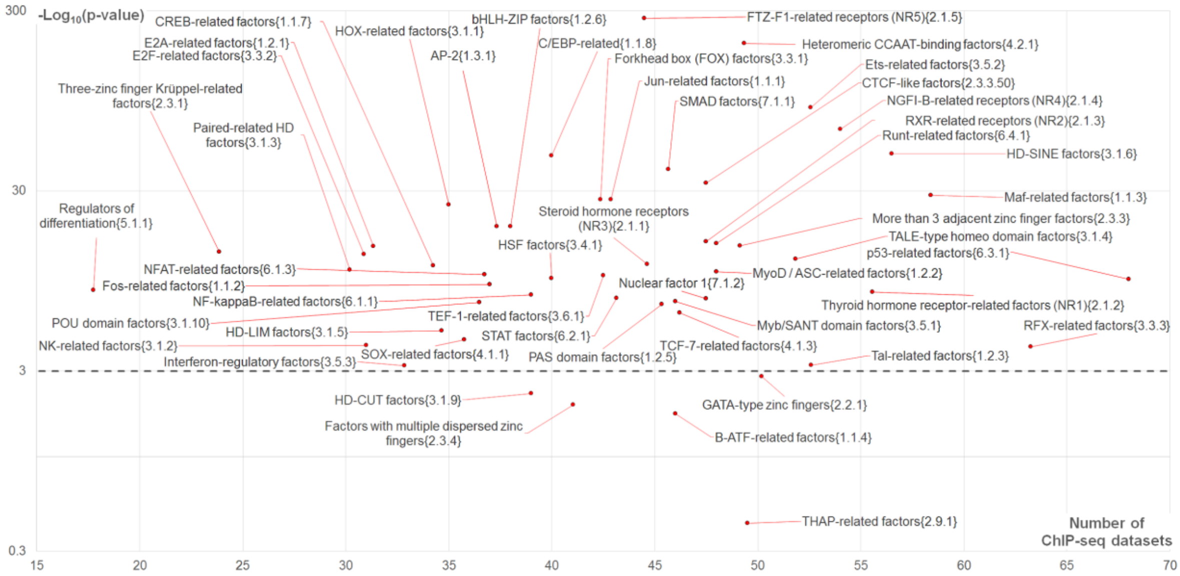


Figure 5. The significance of enrichment of asymmetric CEs toward the partner motifs as a function of CE abundance. The scatterplot shows 50 clades of partner TFs, including 49 families of TFs with at least two motifs and subfamily CTCF-like factors{2.3.3.50} with two motifs from the human core Hocomoco collection [14]. Total number of ChIP-seq datasets is 119 (see Materials and Methods). Axis X implies the number of ChIP-seq dataset with predicted CEs with an overlap of anchor motifs and partner motifs from a specific clade and without taking into account motifs conservation. Axis Y shows the significance of Welch's t-test that for each clade compare the number of ChIP-seq datasets containing asymmetric CEs toward partner motifs and overlaps of motifs and the respective number of datasets containing asymmetric CEs toward anchor motifs. The horizontal dashed line marks Bonferroni's correction for the t-test significance, $-\text{Log}_{10}(\text{p-value}) = 3$.

The majority of clades (26 out of 50) possessed the high significance, p < 1E-10 (Figure 5). The top four clades with the most significant differences were FTZ-F1-related receptors (NR5){2.1.5}, Heteromeric CCAAT-binding factors{4.2.1}, Ets-related factors{3.5.2} and NGFI-B-related receptors (NR4){2.1.4} (p < 1E-273, p < 1E-198, p < 1E-88 and p < 1E-66, respectively). The top twelve clades included also C/EBP-related factors{1.1.8}, CTCF-like factors{2.3.3.50}, Maf-related factors{1.1.3}, Jun-related factors{1.1.1} and Forkhead box (FOX) factors{3.3.1} (Figure 5). The differences were not significant only for five TF families: B-ATF-related factors{1.1.4}, Factors with multiple dispersed zinc fingers{2.3.4}, GATA-type zinc fingers{2.2.1}, HD-CUT factors{3.1.9} and THAP-related factors{2.9.1}.

doi:10.20944/preprints202007.0639.v1

2.4.2. Analysis of top-ranked partner motifs classified according to TFs families

In this section, we are going to perform the detailed analysis of concrete top-ranked partner motifs participating in asymmetric CEs with more conserved partner motifs. This analysis was motivated by occasionally observed imperfect homology of motifs within separate families of TFs (see Materials and Methods), i.e. analysis of previous subsection should be verified by top-ranked predictions for concrete motifs from various top-ranked TF families (Figures 4, 5). Also, we should verify the MCOT with the previous analysis [12] that revealed Jun-like, Ets-like, CTCF-like and THAP11 overrepresented motifs for the fraction of ChIP-seq data lacking canonical motifs of anchor TF.

Thus, at the first step we checked the abundance of partner motifs without taking into account the relationship between conservation of co-occurred motifs. We applied MCOT and selected 30 top-ranked partner motifs from the Hocomoco human core collection [14] that formed CEs with arbitrary relationship between conservation of participant motifs, we excluded homologous pairs anchor-partner (see Materials and Methods) and performed the clustering (Figure 6). Besides the Jun-like and Ets-like motifs, in the list of top-ranked partner motifs we found RFX-like motif, two motifs from Thyroid hormone receptor-related factors (NR1){2.1.2} family, three GATA-like and three p53-like motifs (see Materials and Methods for description of these motifs). We marked in Figure 6 several families of TFs that were mentioned earlier by Worsley Hunt and Wasserman [12], or revealed above in our analysis (Figures 4, 5).

Unfortunately, among the 30 top-ranked motifs (Figure 6) we found many motifs of TFs of the largest two families (More than 3 adjacent zinc finger factors{2.3.3}, Factors with multiple dispersed zinc fingers{2.3.4}, with 76 and 20 partner motifs, respectively; see Table S2). These families belonged to the C2H2 zinc finger TF class with the highest known diversity of DNA binding specificities [21] and respective low specificity in the benchmarking comparison with motifs of other families [22]. Hence, we did not mark motifs of TFs from these two families in Figure 6 and below. Notably, the third largest family ETS-related factors{3.5.2} respected to 19 motifs, for motifs from this family were detected high homology (see Materials and Methods and [23]) and good performance in benchmarking comparisons with motifs of other families [22].

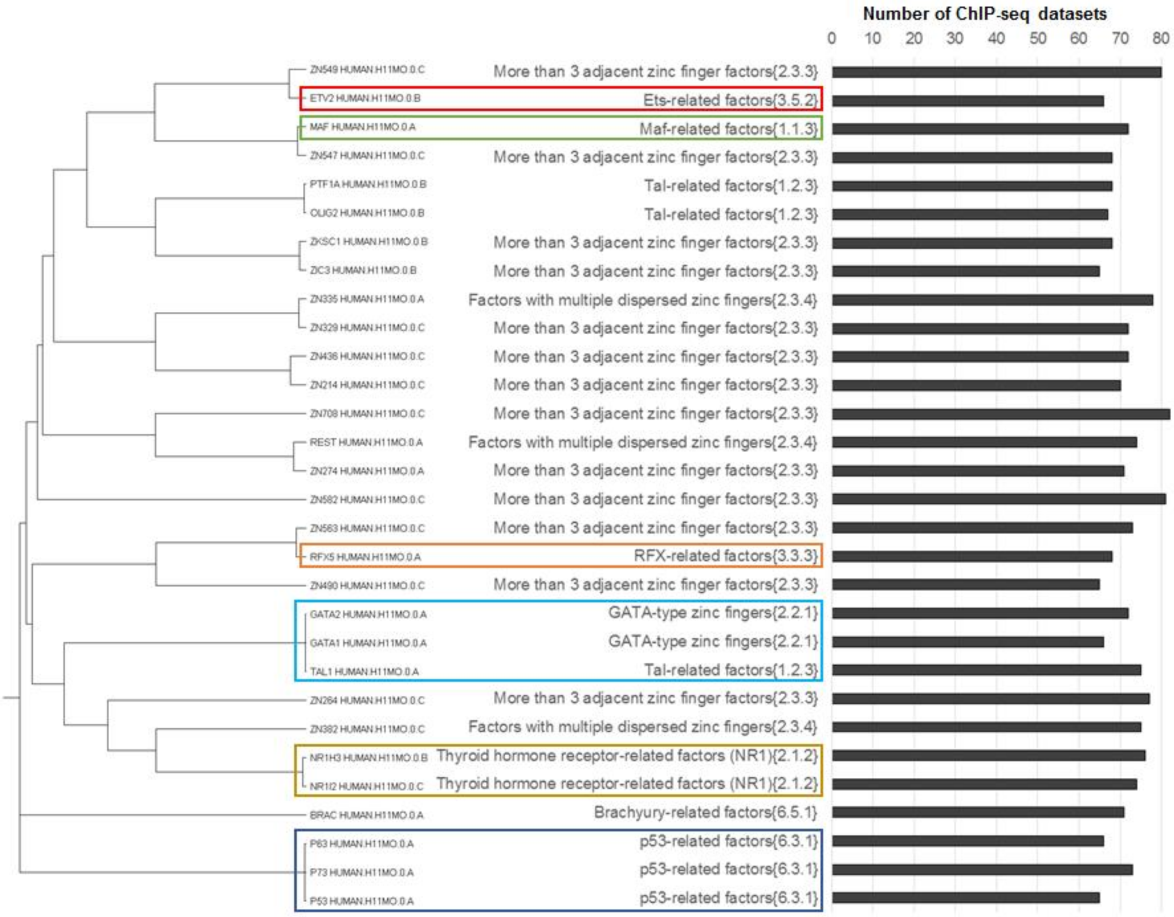


Figure 6. Clustering of 30 top-ranked partner motifs from the Hocomoco human core collection [14], according to their abundance in CEs predicted with an overlap of anchor motifs. We excluded from the analysis CEs containing the significant homology between the anchor and partner motifs. The left/middle/right columns show the tree constructed according to motifs homology, names of TF families [13] and the distribution of the number of ChIP-seq datasets that contained respective CEs. Brown, green, red, orange, blue and aqua boxes mark NR1H3-like motifs from Thyroid hormone receptor-related factors (NR1){2.1.2} family, Jun-like (Maf-related factors{1.1.3}), Ets-like (Ets-related factors{3.5.2}), RFX-like (RFX-related factors{3.3.3}, p53-like (p53-related factors{6.3.1}) and GATA-like (Tal-related factors{1.2.3}) motifs, respectively. Totally, we included in analysis 119 ChIP-seq datasets for human TFs (see Materials and Methods).

Thus, in general the results of our analysis (Figure 6) is in good accordance with our results for 50 TF clades (Figures 4, 5) and with previous results of Worsley Hunt and Wasserman [12]. Thus, Ets-like and Jun-like motifs we found among the top 30, while CTCF-like and THAP11 motifs we did not detected, but still we previously found them among top-ranked TF clades (Figures 4, 5).

Next, we selected 30 top-ranked partner motifs that formed asymmetric CEs toward either anchor or partner motifs, again excluded homologous anchor-partner pairs and performed the clustering (Figure 7). Notably, the separate analysis of asymmetric CEs toward the partner motifs have shown larger variety than that for asymmetric CEs toward the anchor motifs (compare panels A and B of Figure 7). Thus, NR1H3-like, RFX-like, GATA-like motifs and p53-like motifs were found in both lists, Jun-like motifs were absent in both lists. The rank of the best Jun-like motif MAF_HUMAN.H11MO.0.A was only 62 (Table S2).

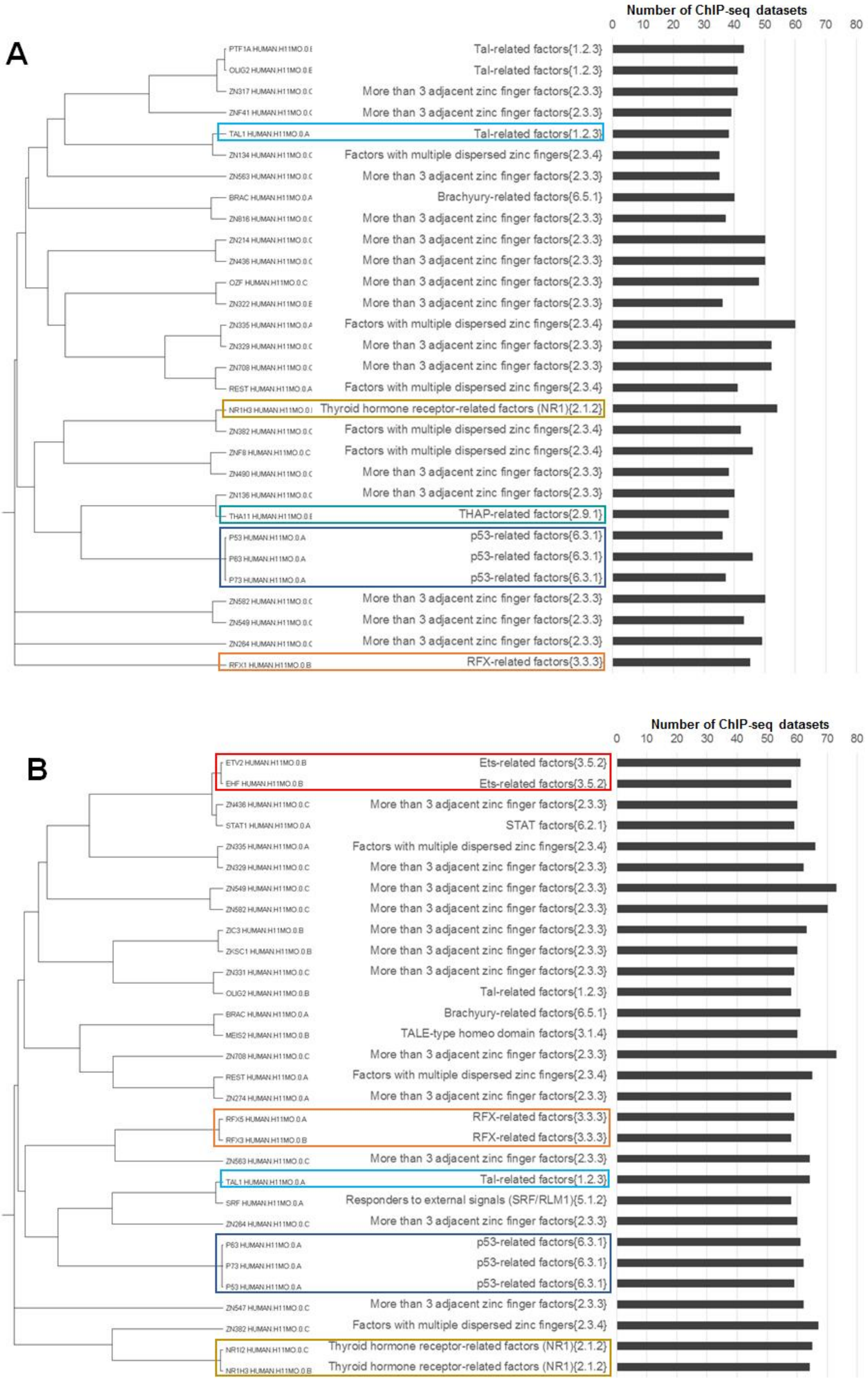

A
Number of ChIP-seq datasets
0
10
20
30
40
50
60
70
80
PTF1A HUMAN.H11MO.0.E Tal-related factors{1.2.3}
OLIG2 HUMAN.H11MO.0.E Tal-related factors{1.2.3}
ZN317 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZNF41 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
TAL1 HUMAN.H11MO.0.A Tal-related factors{1.2.3}
ZN134 HUMAN.H11MO.0.C Factors with multiple dispersed zinc fingers{2.3.4}
ZN563 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
BRAC HUMAN.H11MO.0.A Brachyury-related factors{6.5.1}
ZN816 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZN214 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZN436 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
OZF HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZN322 HUMAN.H11MO.0.E More than 3 adjacent zinc finger factors{2.3.3}
ZN335 HUMAN.H11MO.0.A Factors with multiple dispersed zinc fingers{2.3.4}
ZN329 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZN708 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
REST HUMAN.H11MO.0.A Factors with multiple dispersed zinc fingers{2.3.4}
NR1H3 HUMAN.H11MO.0.I Thyroid hormone receptor-related factors (NR1){2.1.2}
ZN382 HUMAN.H11MO.0.C Factors with multiple dispersed zinc fingers{2.3.4}
ZNF8 HUMAN.H11MO.0.C Factors with multiple dispersed zinc fingers{2.3.4}
ZN490 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZN136 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
THA11 HUMAN.H11MO.0.E THAP-related factors{2.9.1}
P53 HUMAN.H11MO.0.A p53-related factors{6.3.1}
P63 HUMAN.H11MO.0.A p53-related factors{6.3.1}
P73 HUMAN.H11MO.0.A p53-related factors{6.3.1}
ZN582 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZN549 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZN264 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
RFX1 HUMAN.H11MO.0.B RFX-related factors{3.3.3}
B
Number of ChIP-seq datasets
0
10
20
30
40
50
60
70
80
ETV2 HUMAN.H11MO.0.B Ets-related factors{3.5.2}
EHF HUMAN.H11MO.0.B Ets-related factors{3.5.2}
ZN436 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
STAT1 HUMAN.H11MO.0.A STAT factors{6.2.1}
ZN335 HUMAN.H11MO.0.A Factors with multiple dispersed zinc fingers{2.3.4}
ZN329 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZN549 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZN582 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZIC3 HUMAN.H11MO.0.B More than 3 adjacent zinc finger factors{2.3.3}
ZKSC1 HUMAN.H11MO.0.B More than 3 adjacent zinc finger factors{2.3.3}
ZN331 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
OLIG2 HUMAN.H11MO.0.B Tal-related factors{1.2.3}
BRAC HUMAN.H11MO.0.A Brachyury-related factors{6.5.1}
MEIS2 HUMAN.H11MO.0.B TALE-type homeo domain factors{3.1.4}
ZN708 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
REST HUMAN.H11MO.0.A Factors with multiple dispersed zinc fingers{2.3.4}
ZN274 HUMAN.H11MO.0.A More than 3 adjacent zinc finger factors{2.3.3}
RFX5 HUMAN.H11MO.0.A RFX-related factors{3.3.3}
RFX3 HUMAN.H11MO.0.B RFX-related factors{3.3.3}
ZN563 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
TAL1 HUMAN.H11MO.0.A Tal-related factors{1.2.3}
SRF HUMAN.H11MO.0.A Responders to external signals (SRF/RLM1){5.1.2}
ZN264 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
P63 HUMAN.H11MO.0.A p53-related factors{6.3.1}
P73 HUMAN.H11MO.0.A p53-related factors{6.3.1}
P53 HUMAN.H11MO.0.A p53-related factors{6.3.1}
ZN547 HUMAN.H11MO.0.C More than 3 adjacent zinc finger factors{2.3.3}
ZN382 HUMAN.H11MO.0.C Factors with multiple dispersed zinc fingers{2.3.4}
NR1I2 HUMAN.H11MO.0.C Thyroid hormone receptor-related factors (NR1){2.1.2}
NR1H3 HUMAN.H11MO.0.B Thyroid hormone receptor-related factors (NR1){2.1.2}

Figure 7. Clustering of 30 top-ranked partner motifs from the Hocomoco human core collection [14] according to their abundance in CEs predicted with an overlap of anchor motifs. We excluded from the analysis CEs containing the significant homology between the anchor and partner motifs. Panels (A) and (B) show results for CEs with more conserved anchor and partner motifs, respectively. For each panel the left/middle/right columns show the tree constructed according to motifs homology, names of TF families [13] and the distribution of the number of ChIP-seq datasets that contained respective CEs. Brown, green, red, orange, blue, cyan and aqua boxes mark NR1H3-like motifs from Thyroid hormone receptor-related factors (NR1){2.1.2} family, Jun-like (Maf-related factors{1.1.3}), Ets-like (Ets-related factors{3.5.2}), RFX-like (RFX-related factors{3.3.3}, p53-like (p53-related factors{6.3.1}), THAP-related factors{2.9.1} and GATA-like (Tal-related factors{1.2.3}) motifs, respectively. Totally, we included in analysis 119 ChIP-seq datasets for human TFs (see Materials and Methods).

As concerns Jun-like motifs, our analysis that took into account conservation of motifs (Figure 7) seemed to be contradictory to that which did not take it into account (Figure 6). But, previous study [12], that revealed overrepresented Jun-like motifs for the fraction of ChIP-seq data lacking canonical anchor motifs, did not perform the check of homology between the anchor and partner motifs. Hence, we overridden the restriction on the significant homology between CE participants and confirmed that the rank of Jun-like motifs substantially increased (Figure S2). Hence, we presumed that this enrichment of partner Jun-like motifs at least partially was based on their significant similarity to anchor motifs. Thus, we could not confirm the critical importance of Jun-like TFs in cooperative binding with other TFs to DNA.

We may conclude that motifs of TFs from Ets-related factors{3.5.2} family were found only in the list of asymmetric CEs toward partner motifs (Figure 7). Consequently, Ets-like motifs have the most clear among other families tendency to form asymmetric CEs with anchor motifs, so that within these CEs (a) partner motifs have higher conservation than anchor motifs, and (b) the similarity between anchor and partner motifs is absent.

The Full computation flow have shown only two NR1H3-like motifs from the Thyroid hormone receptor-related factors (NR1){2.1.2} family (45 and 30 datasets, Table S2) and three p53-like motifs (33, 30 and 29 datasets; Table S2), these results we computed also for asymmetric CEs toward the partner motifs. In the Partial computation flow we revealed the first-ranked CTCF-like motif from subfamily CTCF-like factors{2.3.3.50}, it detected in only 8 ChIP-seq datasets, while for the Spacer computation flow first three motifs were NFYA-like (Heteromeric CCAAT-binding factors{4.2.1} family) with respecting only 7, 6 and 6 datasets (Table S2).

p53-like, GATA-like and NR1H3-like motifs have shown the enrichment in both cases of asymmetry toward the anchor and partner motifs (Figure 7). Thus, among other families, motifs of ETS family the most clearly demonstrated specific enrichment in asymmetric CEs toward partner motifs. Hence, we may suppose that Ets-like motifs facilitate weak direct interaction of anchor TFs with their cognate binding sites in ChIP-seq peaks. In this case, the ternary complex of anchor TF, TF from ETS family and DNA is formed so that the Ets-like motif is systematically more conserved than anchor motifs, i.e. TFs from ETS family have the leading role in the cooperative interaction with other TFs, when they bind DNA.

3. Discussion

Many studies confirmed that ChIP-seq data possessed a substantial portion of target regions lacking the conserved motifs of respective TFs [12,19]. In the current study, we aimed to clarify whether this portion might respect weak binding motifs of anchor TFs that were located near relatively more conserved motifs of multiple partner TFs. We applied recently developed MCOT package for prediction of motifs co-occurrence with their overlaps and with spacers in a single ChIP-seq dataset [11]. The novelty of our study consisted in analysis of specific CEs with higher conservation of either anchor or partner motifs. We improved the previous algorithm [11] for estimation of the significance of such asymmetric CEs (Figure 1C,D,F) and developed the novel methodology to measure the asymmetry within CEs toward one of participant motifs (Figure 1C,E,G). Next, we have shown the example of significant asymmetry within earlier known CEs FoxA2-HNF1β for ChIP-seq dataset from the liver tissue (Figure 2, [15]). The higher conservation of HNF1β motif in these CEs proposed its leading importance in cooperative binding of both TFs, e.g. presumably HNF1β binding sites were preliminary occupied by HNF1β. This hypothesis is supported by earlier observation, that induced hepatic stem cells were directly induced from mouse embryonic fibroblasts by overexpressing two key TFs, HNF1β and FoxA3 [24]. The next example (Figure 3) illustrated the action of multiple partner motifs co-occurring near the anchor FoxA2 motifs in the same ChIP-seq dataset [15]. We specifically excluded from the analysis peaks with the most conservative and too weak FoxA2 motifs. Peaks with the stringent anchor motifs most probably respect to direct FoxA2 targets, too weak FoxA2 targets potentially required an alternative to PWM model [20], so that an expressive support for FoxA2 binding from partner TFs we expected for intermediate cases of moderately or weakly conserved FoxA2 motifs. Our analysis demonstrated that about 90% of analyzed peaks contained asymmetric pairs of co-occurred anchor and partner motifs, so that partner motifs possessed higher conservation in pairs (Figure 3). Conventionally, in almost all analyses before MCOT, a single threshold for a recognition model of anchor motif was applied, so that weak interactions might be missed by a standard recognition model. Hence, these weakly conserved anchor motifs probably were annotated as indirect or non-specific binding (e.g. in [12,18]). We proposed that in this case multiple overrepresented CEs with higher conservation of partner motifs and lower conservation of anchor motifs, at least partially, explained the absence of canonical motifs of anchor TFs in a substantial portion of a ChIP-seq dataset.

Next, we performed massive analysis with the benchmark ChIP-seq data to study whether partner TFs from specific families possessing common characteristics of DNA-binding domains [13] tended to form specific asymmetric CEs toward partner motifs. As follows from the previous example (Figure 3), such partner TFs might have specific opportunities systematically to mediate the interaction of anchor TFs with their cognate binding sites in ChIP-seq data. Previously, Worsley Hunt and Wasserman [12] for the benchmark ChIP-seq data demonstrated that CTCF-like, Jun-like, Ets-like and THAP11 motifs had overrepresented motifs near summits in peaks lacking the canonical motifs of anchor TFs. These enriched motifs were termed 'zingers' to highlight their outstanding enrichment in ChIP-seq datasets for other studied anchor TFs. With this knowledge, we took our benchmark data of 119 ChIP-seq datasets for 45 distinct TFs (Table S1, [10]) with manually annotated anchor motifs derived from *de novo* motif search [8] and predicted CEs with several additional criteria. In particular, we searched CEs that (a) respected higher conservation of partner motifs than that of anchor motifs, and (b) did not respect the significant similarity between anchor and partner motifs. We proposed that the enrichment of such asymmetric CEs with simultaneous less significant enrichment of the respective CEs with opposite asymmetry, i.e. the higher conservation of anchor motifs, reflected the leading role of partner motifs in cooperative interaction of anchor/partner TF pairs with genomic DNA. Thus, we used similar research strategy as Worsley Hunt and Wasserman [12], but our MCOT algorithm with varied thresholds of both motifs until the very loose (FPR = 5E-4) allowed us deduce potential CEs with almost imperceptible with a canonical threshold occurrences of anchor motifs. Moreover, out tool had advantage for analysis of co-occurrence of motifs with an overlap, that have been missed in previous studies for a single ChIP-seq dataset [8,9,24,25]. Additionally, conventionally applied masking procedure (e.g. in [12]) for anchor motifs inevitably destroyed overlapping partner motifs, though overlapping of motifs were observed notably higher than their co-occurrence with a spacer [10,27,11].

Our results substantially extended and supplemented the previous study [12] (Figures 4-7). We confirmed their conclusions concerning CTCF-like, Jun-like, Ets-like and THAP11 motifs. However, our analysis brought many details concerning specific families of TFs. Thus, we explained the enrichment of Jun-like motifs by their similarity to anchor motifs (Figure S2, Figure 7). Also, we found partner motifs of TFs from THAP-related factors{2.9.1} among top-ranked in the list respecting CEs with arbitrary conservation of motifs (Figure 6) and in the list respecting to asymmetric CEs toward anchor motifs (Figure 7A). Moreover, the THAP-related factors{2.9.1} family was detected among only five families among total 50 clades that did not possess the significant enrichment of the abundance of asymmetric CEs toward partner motifs vs. that for asymmetric CEs toward anchor motifs (Figure 5).

The detailed analysis (Figure 4,5,7) demonstrated that besides proposed earlier [12] CTCF-like, Jun-like, Ets-like and THAP11 motifs, other motifs, in particular NR1H3-like, RFX-like, p53-like, NFYA-like and GATA-like also systematically promoted binding of anchor TFs in ChIP-seq data. We may conclude that Ets-like motifs comprised CEs with their highest conservation relative to anchor motifs, respective CEs were not enriched in the list of top ranked predictions for asymmetry toward the anchor motifs, and ETS-like motifs were not significantly similar to the anchor motifs participating in significantly enriched CEs.

We presumed that the function of partner TFs did not consist in only indirect binding of anchor TFs ('tethering'); rather, the more conserved motifs of partner TFs overlapped less conserved motifs of anchor TFs. We propose the 'permanent' model of cooperative binding of anchor and partner TFs (Figure 8). If an anchor TF binds genomic DNA directly, then the respective anchor motif is strongly conserved (Figure 8A). The appearance of another TF (partner) may induce the protein-protein interaction anchor-partner that transforms this direct binding of anchor TF to CE anchor-partner with more or less conserved anchor motif (Figure 8B,C), so that the anchor motif becomes moderately or weakly conserved, respectively. Finally, the anchor motif loses even a weak contact with DNA, so that we may find in DNA only the motif of partner TF (Figure 8D).

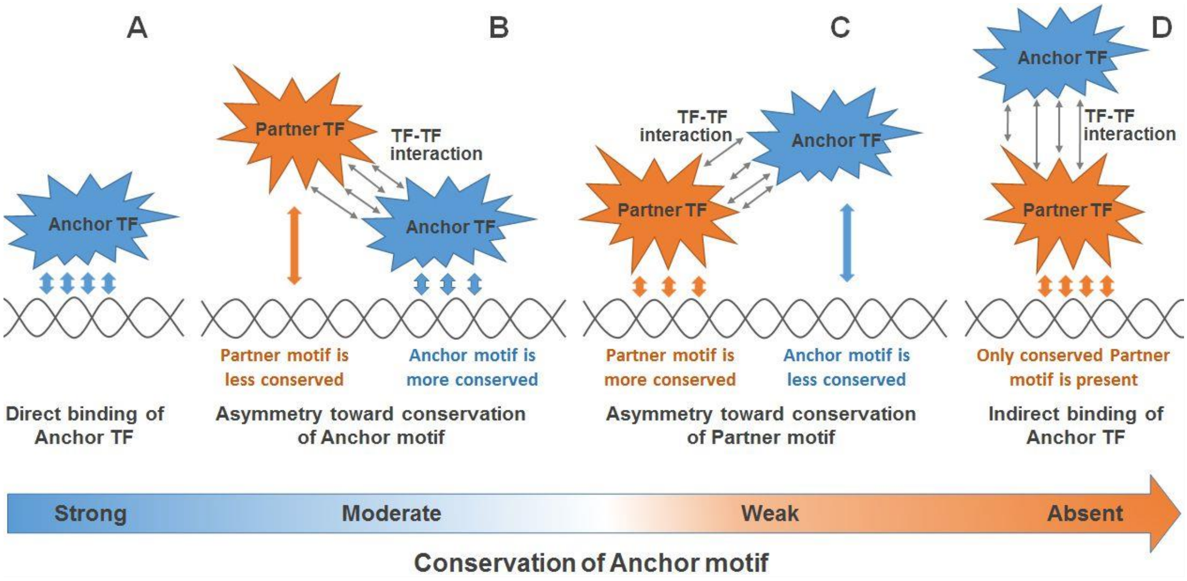

A
B
C
D
Anchor TF
Partner TF
TF-TF interaction
Anchor TF
TF-TF interaction
Partner TF
Anchor TF
Anchor TF
TF-TF interaction
Partner TF
Partner motif is less conserved
Anchor motif is more conserved
Partner motif is more conserved
Anchor motif is less conserved
Only conserved Partner motif is present
Direct binding of Anchor TF
Asymmetry toward conservation of Anchor motif
Asymmetry toward conservation of Partner motif
Indirect binding of Anchor TF
Strong
Moderate
Weak
Absent
Conservation of Anchor motif

Figure 8. The 'permanent' model of cooperative binding of anchor and partner TFs for explanation of a substantial portion of ChIP-seq data lacking conserved motifs of anchor TFs. Panel (A) respects to the most conserved motifs of an anchor TF in a ChIP-seq dataset, such motifs are in most cases overrepresented and successively recognized as the canonical motif of anchor TF. However, an anchor TF often participates in TF-TF interactions with multiple partner TFs. Thus, a whole conservation of anchor-partner CE is subdivided between anchor and partner motifs. We propose here two options: the anchor motif preserves the higher conservation than the partner motif (B), or the anchor motif has less conserved motif than the partner motif (C). Finally, an anchor TF binds to DNA indirectly (D), e.g. if the heterodimer of anchor/partner TFs binds with DNA only through partner TF. The long arrow in the bottom reflects the permanent decrease/increase of the conservation of the anchor/partner motif. The numbers of red/blue arrows between each TF and DNA reflect the conservation of respective motif.

The family of ETS-related TFs consisted of 27-28 members, which were further classified into several subfamilies [13,21,23]. According to the comparative analysis of human TFs [21], besides the ETS family, only several other TF families or superfamilies, e.g. Nuclear receptors, STAT and T-box, had the complete coverage of known motifs and absence of secondary motifs.

Recent all-against-all benchmarking of PWM models [22] suggested that the majority of ETS members have indistinguishable DNA binding specificity according to in vitro HT-SELEX assays. Thus, while a single PWM for ELK1 (MA0028.2 from JASPAR) was the best predictor for multiple TFs from the ETS family for in vivo and in vitro experiments; this matrix also was the best performer for ChIP-seq in vivo experiments for ten TFs, only five of which were ETS family members. For the rest five unrelated TFs it was proposed the recruitment to their target binding sites through protein-protein interactions with a DNA-bound ETS factor. This hypothesis is in excellence accordance with our analysis (Figures 4, 5).

The previous analysis of genome binding of ETS family members [23] proposed that DNA-binding specificity differences alone could not explain genomic binding diversity of TFs from ETS family. Authors proposed two possible mechanisms to achieve specificity for a certain family member: the divergent expression patterns of various family members and the cooperative binding of ETS factors with other TFs. The first mechanism was at least partially supported by (a) only partial overlapping of expression patterns of various family members revealed in transcriptome data [28] and (b) knock-down experiments replacing one member to another [29,30]. Results of our study as well as the previous reviews on protein-protein interaction of ETS TFs [31-33] strongly supported the second mechanism, i.e. combinatorial control of transcription as a characteristic property of ETS family members.

Outstanding properties of TFs from ETS family were also supported by protein structure analysis. In contrast to prokaryotes, the majority of eukaryotic TFs contained long stretches of intrinsically disordered regions (IDRs), which were sequences that did not adopt a stably structured conformation but they were essential for activity [34]. In TFs, IDRs were highly enriched around DNA binding domains (DBDs), which displayed electrostatically biased surfaces to their surroundings [35]. In the ETS family IDRs and highly stable α-helices flanking the DBD (ETS domain) were autoinhibitory for ETS1, ETS2, ETV6, ERG, and ETV1/4/5 binding to DNA [32,36-39]. ETS1, SPI1 and some other members of ETS family were also regulated by another IDR serine-rich region. DBD was autoinhibited in several family members by different mechanisms. Thus, a serine-rich IDR allosterically inhibited DNA binding of ETS1 through phosphorylation-enhanced interactions with the structured DBD and flanking N- and C-terminal inhibitory α-helices [40,41], or a single flanking C-terminal α-helix sterically inhibited DNA binding of ETV6 [42,36,43]. For ETV4 acetylation of selected lysines within the N-terminal IDR activated DNA binding, a C-terminal α-helix perturbed the conformation of its DNA-recognition helix [39]. Recently, experimental study of relatively distant paralogous ETS family members ETS1 and SPI1 have shown that the binding of DNA and the synthetic peptides containing IDRs by the DBD were mutually exclusive [44].

Thus, subfamily-specific α-helices that flank DBD and TF partners through IDRs could modify during TF-TF interaction the equilibrium between active and inactive states; and also post-translational modifications within IDRs specifically regulated an individual ETS factor [39]. Hence, the regulatory strategy of TFs from ETS family consisted in activation through recruitment by other coactivators [45]. This conclusion is in good accordance with the results of our study.

4. Materials and Methods

4.1. MCOT: classification of co-occurred motifs

In the current study, we applied the MCOT package as described earlier [11] with some improvement (see below). This tool annotated pairs of overrepresented motifs, i.e. CEs. Input data of tool compiled peaks of ChIP-seq dataset in Fasta format, the anchor motif (nucleotide frequency matrix) that respected to potential BSs of target TF, and either a partner motif or the list of partner motifs extracted from Hocomoco human or mouse core collections [14].

4.2. Composite elements search and annotation

MCOT classified CEs according to mutual orientation of motifs, e.g. for heterotypic CEs there were four distinct orientations (Figure 1A). There were three distinct cases of mutual locations: Full/Partial overlaps and Spacer, consequently MCOT used five computation flows (Full, Partial, Overlap, Spacer and Any, Figure 1B). MCOT applied the recognition model of PWM for mapping motifs in peaks. For each matrix, five thresholds {T_1, ..., T_5} were used according to the unified set of expected FPRs for a whole-genome dataset of promoters, {5.24E-5, 1.02E-04, 1.9E-4, 3.33E-4, 5E-4}. The profile of the most stringent hits contained PWM scores $T \geq T_1$, the next profile comprised scores in the range $T_2 \geq T > T_1$, etc. Hence, MCOT subdivided all CEs into two classes: those with more conservative anchor and partner motifs (Figure 1C). The conservation of motif hit we estimated through the expected FPR as $-\text{Log}_{10}$(FPR). For each of 5x5 = 25 combinations of motifs conservation and each computation flow MCOT compiled the 2x2 contingency table (Figure 1F) and computed the significance of Fisher's exact test that compared the abundance of CEs in peaks with that for the background model. The background dataset was generated as described earlier [11].

4.3. Significance of asymmetry within composite elements

We improved the original MCOT algorithm [11] to calculate the asymmetry within CE as follows. For each CE we estimated the conservation of motifs with $-\text{Log}_{10}$[FPR] value. Than we applied the criteria

- {$-\text{Log}_{10}$[FPR(Anchor)] > $-\text{Log}_{10}$[FPR(Partner)]} and
- {$-\text{Log}_{10}$[FPR(Anchor)] ≤ $-\text{Log}_{10}$[FPR(Partner)]}

and classified all predicted CEs into two classes with more conservative anchor or partner motifs. Next, for each class we computed the enrichment p-value that compared fractions of peaks containing/not containing CEs in foreground and background datasets (Figure 1F). To estimate the asymmetry of CE we applied the Fisher's exact test that compared the portions of CEs with more conserved anchor and partners motifs in foreground and background datasets (Figure 1**G**). We assigned to the asymmetry significance $-\text{Log}_{10}$[P-value] the sign '+' in the case of enrichment toward the anchor motif, otherwise, sign '-' denoted the enrichment toward the partner motif. The conservation of each motif we estimated by the expectation of it occurrence in whole-genome promoter dataset with $-\text{Log}_{10}$(FPR) measure. Next, for foreground and background datasets of sequences we compiled the full lists of predicted CEs (see notation Obs and Exp, respectively, in Figure 1**F**,**G**). We classified the conservation of each motif within the ranges of twelve conservation levels as follows [<3.5], [3.5..3.7], [3.7..3.9] etc. up to [5.3..5.5] and [>5.5]. We computed the counts of CEs from foreground and background datasets $\text{Obs}_{i,j}$ and $\text{Exp}_{i,j}$ that had distinct combinations of conservation levels. Here indices *i* and *j* denote conservation levels for anchor and partner motifs. Finally, the per mille measure transforms the absolute CE counts to relative ones as follow: {$1000*\text{Obs}_{i,j}$/Obs)} and {$1000*\text{Exp}_{i,j}$/Exp}.

4.4. Bonferroni correction for significance

To take into account multiple comparisons we applied the Bonferroni's correction and used the following critical values for (a) the significance of CEs without taking into account motifs conservation, (b) the significance of asymmetric CEs toward either anchor or partner motifs and (c) CE asymmetry: $0.05/(N_{FOR}*N_{BACK}*N_{FLOW}*N_{THR}*N_{THR})$, $0.05/(N_{FOR}*N_{BACK}*N_{FLOW}*2)$ and $0.05/(N_{FOR}*N_{BACK}*N_{FLOW})$, respectively. Here N_{FOR} and N_{BACK} means the size of foreground and background datasets (these numbers imply the number of peaks and random sequences, which generated in MCOT [11], $N_{FLOW} = 5$ designates the number of MCOT computation flows and $N_{THR} = 5$ means the number of thresholds for each motif.

4.5. Massive analysis of ChIP-seq data

In the current study, we complemented previously published benchmark ChIP-seq data [11] for human TFs, so the whole collection consisted of 119 ChIP-seq datasets for 45 TFs (Table S1). As in earlier study [11], for each dataset we annotated the results of *de novo* motif search [8], manually selected enriched motifs respecting to the anchor TF and approved the homology between the *de novo* detected and known motifs [46]. We applied the MCOT as described earlier and above in this study (https://gitlab.sysbio.cytogen.ru/academiq/mcot-kernel, [11]). In particular, 396 partner motifs of human TFs as described earlier were extracted from the Hocomoco human core collection [14]. We used the MEGA package to draw trees that showed the similarity of motifs (https://www.megasoftware.net/, [47]. We used the classification of human and mouse TFs according to the characteristics of their DNA-binding domains (http://tfclass.bioinf.med.uni-goettingen.de/, [13]). We supplied all partner motifs with the names of respective families and classified all motifs into 67 distinct families of TFs. Since the consequent analysis was based on the recognition of motifs, we performed the pairwise comparison of homology of all partner motifs with the motif comparison tool from MCOT [11] ($p<0.05$ for at least one of two motifs similarity measures). In our analysis we preserved the classification of motifs according to their families [13], but annotated together homologous motifs from various families. In particular, according to previous data [12] we distinguished following groups of motifs:

- Jun-like, out total 18 motifs of Jun-related{1.1.1}, Fos-related{1.1.2} and Maf-related{1.1.3} families 15 were homologous;
- Ets-like, out of total 19 motifs of Ets-related factors{3.5.2} family 14 were homologous;
- CTCF-like, two homologous motifs constituted the subfamily CTCF-like factors{2.3.3.50} of the largest family More than 3 adjacent zinc finger factors{2.3.3} consisting of 76 motifs;
- two non-homologous motifs THA11_HUMAN.H11MO.0.B and THAP1_HUMAN.H11MO.0.C constituted THAP-related factors{2.9.1} family.

Moreover, in our analysis we also paid attention to following motifs, classified according to TF families:

- p53-like, all 3 motifs from family p53-related factors{6.3.1} were homologous;
- RFX-like, all 4 motifs from family RFX-related factors{3.3.3} were homologous;
- GATA-like, all 5 motifs from family GATA-type zinc fingers{2.2.1} were homologous, we added to them their homologue TAL1_HUMAN.H11MO.0.A from Tal-related factors{1.2.3} family (rest participants of this family were not homologous to GATA-like motifs);
- NR1H3-like motifs, 4 motifs from Thyroid hormone receptor-related factors (NR1){2.1.2} family (NR1H3_HUMAN.H11MO.0.B, THA_HUMAN.H11MO.0.C, NR1I3_HUMAN.H11MO.0.C, NR1I2_HUMAN.H11MO.0.C) were homologous, this family consisted of 14 motifs; NR1H3-like motifs had close homologous motifs in families of Steroid hormone receptors (NR3){2.1.1} (e.g. ERR1_HUMAN.H11MO.0.A) and RXR-related receptors (NR2){2.1.3} (e.g. COT2_HUMAN.H11MO.0.A);

- NFYA-like, all 3 motifs from family Heteromeric CCAAT-binding factors{4.2.1} were homologous.

We selected for consequent analysis 49 families with at least two motifs among all 67 families respecting all 396 partner motifs. We also included in analysis the subfamily CTCF-like factors{2.3.3.50}, since CTCF-like motifs previously were annotated [12]. Thus, we included in analysis 50 clades of partner TFs, including 49 families and one subfamily.

We performed prediction of potential CEs with the MCOT for the benchmark data of 119 ChIP-seq datasets (see Table S1). We proposed that homology between the anchor and partner motifs might influence partner motifs enrichment. Hence, we excluded CEs consisting of significantly similar partner and anchor motifs. The significant similarity we detected with at least one of two motifs similarity measures ($p < 0.05$, [11]. We applied Bonferroni's correction for the significance of CEs (see above) and counted ChIP-seq datasets with significant CEs separately for five MCOT computation flows.

5. Conclusions

We proposed the approach for computation of the significance of co-occurrence of asymmetric CEs anchor-partner with one of participant motifs more conservative than another one, and for asymmetry within pairs of co-occurred motifs; we applied our approach for motifs of partner TFs from various families overrepresented near motifs of anchor TFs in ChIP-seq data. We demonstrated that for partner motifs of almost all families of TFs only for overlapping pairs of anchor and partner motifs but not for pairs with a spacer, pairs with higher conservation of partner motifs were significantly more abundant than those with higher conservation of anchor motifs. This observation explained a substantial portion of ChIP-seq data lacking even moderately conserved anchor motifs. We found that the asymmetric CEs toward partner motifs were the most reliable for partner motifs of TFs from ETS family. Hence, motifs of TFs from ETS family tended to mediate interaction of anchor TFs with genomic DNA.

Supplementary Materials: Supplementary materials can be found at www.mdpi.com/xxx/s1.

Author Contributions: Conceptualization V.L.; Methodology V.L., D.O.; Software V.L.; Investigation V.L., Writing—original draft preparation V.L.; Writing—review and editing, V.L., E.Z. and T.M. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by Russian Foundation for Basic Research Project # 18-29-13040 and State Budget Project # 0324–2019-0040-C-01.

Acknowledgments: The bioinformatics data analysis was performed in part on the equipment of the Bioinformatics Shared Access Center with the support of State Budgeted Project # 0324-2019-0040-C-01.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

BS	Binding Site
CE	Composite Element
DBD	DNA Binding Domain
FPR	False Positive Rate
IDR	Intrinsically Disordered Region
MCOT	Motifs Co-Occurrence Tool
PWM	Position Weight Matrix

References

1. Morgunova, E.; Taipale, J. Structural perspective of cooperative transcription factor binding. *Curr. Opin. Struct. Biol.* **2017**, *47*, 1–8. https://doi.org/10.1016/j.sbi.2017.03.006

2. Reiter, F.; Wienerroither, S., Stark, A. Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.* **2017,** *43*, 73–81. https://doi.org/10.1016/j.gde.2016.12.007
3. Mayran, A.; Drouin, J. Pioneer transcription factors shape the epigenetic landscape. *J Biol Chem.* **2018,** *293*, 13795-13804. https://doi.org/10.1074/jbc.R117.001232
4. Lai, X.; Verhage, L.; Hugouvieux, V.; Zubieta, C. Pioneer factors in animals and plants-colonizing chromatin for gene regulation. *Molecules* **2018**, *23*, e1914. https://doi.org/10.3390/molecules23081914
5. Zaret, K.S.; Carroll, J.S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **2011**, *25*, 2227-2241. https://doi.org/10.1101/gad.176826.111
6. Nagy, G.; Nagy, L. Motif grammar: the basis of the language of gene expression. *Comput Struct Biotec.* (in press), doi: https://doi.org/10.1016/j.csbj.2020.07.007
7. Lloyd, S.M.; Bao, X. Pinpointing the genomic localizations of chromatin-associated proteins: the yesterday, today, and tomorrow of ChIP-seq. *Curr Protoc Cell Biol.* **2019**, *84*, e89. https://doi.org/10.1002/cpcb.89
8. Heinz, S.; Benner, C.; Spann, N.; Bertolino, E.; Lin, Y.C.; Laslo, P.; Cheng, J.X.; Murre, C.; Singh, H.; Glass, C.K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* **2010**, *38*, 576-589. https://doi.org/10.1016/j.molcel.2010.05.004
9. Whitington, T.; Frith, M.C.; Johnson, J.; Bailey, T.L. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res.* **2011**, *39*, 98. https://doi.org/10.1093/nar/gkr341
10. Jankowski, A.; Prabhakar, S.; Tiuryn, J. TACO: a general-purpose tool for predicting cell-type-specific transcription factor dimers. *BMC Genomics* **2014**, *15*, 208. https://doi.org/10.1186/1471-2164-15-208
11. Levitsky, V.; Zemlyanskaya, E.; Oshchepkov, D.; Podkolodnaya, O.; Ignatieva, E.; Grosse, I.; Mironova, V.; Merkulova, T. A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. *Nucleic Acids Res.* **2019**, *47*, e139. https://doi.org/10.1093/nar/gkz800
12. Worsley Hunt, R.; Wasserman, W.W. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.* **2014**, *15*, 412. https://doi.org/10.1186/s13059-014-0412-4
13. Wingender, E.; Schoeps, T.; Haubrock, M.; Krull, M.; Dönitz, J. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.* **2018,** *46*, D343-D347. https://doi.org/10.1093/nar/gkx987
14. Kulakovskiy, I.V.; Vorontsov, I.E.; Yevshin, I.S.; Sharipov, R.N.; Fedorova, A.D.; Rumynskiy, E.I.; Medvedeva, Y.A.; Magana-Mora, A.; Bajic, V.B.; Papatsenko, D.A.; et al. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.* **2018**, *46*, D252-D259. https://doi.org/10.1093/nar/gkv1249
15. Wederell, E.D.; Bilenky, M.; Cullum, R.; Thiessen, N.; Dagpinar, M.; Delaney, A.; Varhol, R.; Zhao, Y.; Zeng, T.; Bernier, B.; et al. Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.* **2008**, *36*, 4549–4564. https://doi.org/10.1093/nar/gkn382
16. Wallerman, O.; Motallebipour, M.; Enroth, S.; Patra, K.; Bysani, M.S.; Komorowski, J.; Wadelius, C. Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Res.* **2009,** *37*, 7498–7508. https://doi.org/10.1093/nar/gkp823

17. Wang, D.; Garcia-Bassets, I.; Benner, C.; Li, W.; Su, X.; Zhou, Y.; Qiu, J.; Liu, W.; Kaikkonen, M.U; Ohgi, K.A.; et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **2011**, *474*, 390–394. https://doi.org/10.1038/nature10006
18. Tsankov, A.M.; Gu, H.; Akopian, V.; Ziller, M.J.; Donaghey, J.; Amit, I.; Gnirke, A.; Meissner, A. Transcription factor binding dynamics during human ES cell differentiation. *Nature* **2015**, *518*, 344–349. https://doi.org/10.1038/nature14233
19. Gheorghe, M.; Sandve, G.K.; Khan, A.; Chèneby, J.; Ballester, B.; Mathelier, A. A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res.* **2019**, *47*, e21. https://doi.org/10.1093/nar/gky1210
20. Levitsky, V.G.; Kulakovskiy, I.V.; Ershov, N.I.; Oshchepkov, D.Y.; Makeev, V.J.; Hodgman, T.C.; Merkulova, T.I. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genomics* **2014,** *15*, 80. https://doi.org/10.1186/1471-2164-15-80
21. Lambert, S.A.; Jolma, A.; Campitelli, L.F.; Das, P.K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T.R.; Weirauch, M.T. The Human transcription factors. *Cell* **2018**, *172*, 650–665. https://doi.org/10.1016/j.cell.2018.01.029
22. Ambrosini, G.; Vorontsov, I.; Penzar, D.; Groux, R.; Fornes, O.; Nikolaeva, D. D.; Ballester, B.; Grau, J.; Grosse, I.; Makeev, V.; Kulakovskiy, I.; Bucher, P. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biol* **2020**, *21*, 114. https://doi.org/10.1186/s13059-020-01996-3
23. Wei, G.H.; Badis, G.; Berger, M.F.; Kivioja, T.; Palin, K.; Enge, M.; Bonke, M.; Jolma, A.; Varjosalo, M.; Gehrke, A.R.; et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.* **2010**, *29*, 2147–2160. https://doi.org/10.1038/emboj.2010.106
24. Yu, B.; He, Z.Y.; You, P.; Han, Q.W.; Xiang, D.; Chen, F.; Wang, M.J.; Liu, C.C.; Lin, X.W.; Borjigin, U.; et al. Reprogramming fibroblasts into bipotential hepatic stem cells by defined factors. *Cell Stem Cell.* **2013**, *13*, 328–340. https://doi.org/10.1016/j.stem.2013.06.017
25. Guo, Y.; Mahony, S.; Gifford, D.K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol.* **2012**, *8*, e1002638. https://doi.org/10.1371/journal.pcbi.1002638
26. Kazemian, M.; Pham, H.; Wolfe, S.A.; Brodsky, M.H.; Sinha, S. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. *Nucleic Acids Res.* **2013,** *41*, 8237-8352. https://doi.org/10.1093/nar/gkt598
27. Jankowski, A.; Szczurek, E.; Jauch, R.; Tiuryn, J.; Prabhakar, S. Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers. *Genome Res.* **2013**, *23*, 1307-1318. https://doi.org/10.1101/gr.154922.113
28. Richardson, L., Venkataraman, S., Stevenson, P., Yang, Y., Burton, N., Rao, J., Fisher, M., Baldock, R.A., Davidson, D.R., Christiansen, J.H. EMAGE mouse embryo spatial gene expression database: 2010 update. *Nucleic Acids Res.* **2010**, *38*, D703-D709. https://doi.org/10.1093/nar/gkp763
29. Dahl, R., Ramirez-Bergeron, D.L., Rao, S., Simon, M.C. Spi-B can functionally replace PU.1 in myeloid but not lymphoid development. *EMBO J.* **2002**, *21*, 2220–2230. https://doi.org/10.1093/emboj/21.9.2220

30. DeKoter, R.P., Lee, H J., Singh, H. PU.1 regulates expression of the interleukin-7 receptor in lymphoid progenitors. *Immunity* **2002**, *16*, 297–309. https://doi.org/10.1016/s1074-7613(02)00269-8

31. Verger, A.; Duterque-Coquillaud, M. When Ets transcription factors meet their partners. *BioEssays* **2002**, *24*, 362–370. https://doi.org/10.1002/bies.10068

32. Hollenhorst, P.C.; McIntosh, L.P.; Graves, B.J. Genomic and biochemical insights into the specificity of ETS transcription factors. *Annu Rev Biochem.* **2011**, *80*, 437–471.
https://doi.org/10.1146/annurev.biochem.79.081507.103945

33. Cooper, C.D., Newman, J.A., Gileadi, O. Recent advances in the structural molecular biology of Ets transcription factors: interactions, interfaces and inhibition. *Biochem Soc Trans.* **2014**, *42*, 130-138. https://doi.org/10.1042/BST20130227

34. Liu, J.; Perumal, N.B.; Oldfield, C.J.; Su, E.W.; Uversky, V N.; Dunker, A.K. Intrinsic disorder in transcription factors. *Biochemistry* **2006**, *45*, 6873-6888. https://doi.org/10.1021/bi0602718

35. Guo, X.; Bulyk, M. L.; Hartemink, A. J. Intrinsic disorder within and flanking the DNA-binding domains of human transcription factors, in *Proceedings of the Pacific Symposium on Biocomputing 2012*, Kohala Coast, Hawaii, USA, 3–7 January 2012; Altman, R.B., Dunker, A.K., Hunter, L., Murray, T., Klein, T.E., Eds.; Publisher: World Scientific, Singapore, 2011, pp. 104–115. https://doi.org/10.1142/9789814366496_0011

36. Coyne, H.J.; 3rd, De, S.; Okon, M.; Green, S.M.; Bhachech, N.; Graves, B J.; McIntosh, L.P. Autoinhibition of ETV6 (TEL) DNA binding: appended helices sterically block the ETS domain. *J Mol Biol.* **2012**, *421*, 67–84. https://doi.org/10.1016/j.jmb.2012.05.010

37. Regan, M.C.; Horanyi, P.S.; Pryor, E.E.; Jr, Sarver, J.L.; Cafiso, D.S.; Bushweller, J.H. Structural and dynamic studies of the transcription factor ERG reveal DNA binding is allosterically autoinhibited, *Proc Natl Acad Sci U S A* **2013**, *110*, 13374–13379. https://doi.org/10.1073/pnas.1301726110

38. Newman, J. A.; Cooper, C. D.; Aitkenhead, H.; Gileadi, O. Structural insights into the autoregulation and cooperativity of the human transcription factor ETS-2. *J Biol Chem.* **2015**, *290*, 8539–8549. https://doi.org/10.1074/jbc.M114.619270

39. Currie, S.L.; Lau, D.; Doane, J.J.; Whitby, F.G.; Okon, M.; McIntosh, L.P.; Graves, B.J.; Structured and disordered regions cooperatively mediate DNA-binding autoinhibition of ETS factors ETV1, ETV4 and ETV5. *Nucleic Acids Res.*; **2017,** *45*, 2223–2241. https://doi.org/10.1093/nar/gkx068

40. Lee, G.M.; Donaldson, L.W.; Pufall, M.A.; Kang, H.S.; Pot, I.; Graves, B.J.; McIntosh, L.P. The structural and dynamic basis of Ets-1 DNA binding autoinhibition. *J Biol Chem.* **2005**, 280:7088–7099. https://doi.org/10.1074/jbc.m410722200

41. Pufall, M.A.; Lee, G.M.; Nelson, M.L.; Kang, H.S.; Velyvis, A.; Kay, L.E.; McIntosh, L.P.; Graves, B.J. Variable control of Ets-1 DNA binding by multiple phosphates in an unstructured region. *Science* **2005**, *309*,142–145. https://doi.org/10.1126/science.1111915

42. Green, S.M.; Coyne, H.J. 3rd, McIntosh, L.P.; Graves, B.J. DNA binding by the ETS protein TEL (ETV6) is regulated by autoinhibition and self-association. *J Biol Chem.* **2010**, *285*, 18496–18504. https://doi.org/10.1074/jbc.m109.096958

doi:10.20944/preprints202007.0639.v1

43. De, S.; Chan, A.C.; Coyne, H.J. 3rd, Bhachech, N.; Hermsdorf, U.; Okon, M.; Murphy, M.E.; Graves, B.J.; McIntosh, L.P. Steric mechanism of auto-inhibitory regulation of specific and non-specific DNA binding by the ETS transcriptional repressor ETV6. *J Mol Biol.* **2014**, *426,* 1390–1406. https://doi.org/10.1016/j.jmb.2013.11.031

44. Perez-Borrajero, C., Lin, C.S., Okon, M., Scheu, K., Graves, B.J., Murphy, M., McIntosh, L.P. The biophysical basis for phosphorylation-enhanced DNA-binding autoinhibition of the ETS1 transcription factor. *J Mol Biol.* **2019**, *431*, 593-614. https://doi.org/10.1016/j.jmb.2018.12.011

45. Xhani, S.; Lee, S.; Kim, H. M.; Wang, S.; Esaki, S.; Ha, V.; Khanezarrin, M.; Fernandez, G.L.; Albrecht, A. V.; Aramini, J. M.; Germann, M.W.; Poon, G. Intrinsic disorder controls two functionally distinct dimers of the master transcription factor PU.1. *Sci Adv.* **2020**, *6*, eaay3178. https://doi.org/10.1126/sciadv.aay3178

46. Gupta, S.; Stamatoyannopolous, J.A.; Bailey, T.L.; Noble, W.S. Quantifying similarity between motifs. *Genome Biol.* **2007**, *8*, R24. https://doi.org/10.1186/gb-2007-8-2-r24

47. Kumar, S.; Stecher, G.; Li,M.; Knyaz, C.; Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. **2018**, *35*, 1547-1549. https://doi.org/10.1093/molbev/msy096