International Journal of
Molecular Sciences

MDPI

Article

Peak Scores Significantly Depend on the Relationships between Contextual Signals in ChIP-Seq Peaks

Oleg V. Vishnevsky [1,2,*], Andrey V. Bocharnikov [2] and Elena V. Ignatieva [1,2]

[1] Institute of Cytology and Genetics, 630090 Novosibirsk, Russia; eignat@bionet.nsc.ru
[2] Department of Natural Science, Novosibirsk State University, 630090 Novosibirsk, Russia; andrey.bocharnikov@gmail.com
* Correspondence: oleg@bionet.nsc.ru

Abstract: Chromatin immunoprecipitation followed by massively parallel DNA sequencing (ChIP-seq) is a central genome-wide method for in vivo analyses of DNA-protein interactions in various cellular conditions. Numerous studies have demonstrated the complex contextual organization of ChIP-seq peak sequences and the presence of binding sites for transcription factors in them. We assessed the dependence of the ChIP-seq peak score on the presence of different contextual signals in the peak sequences by analyzing these sequences from several ChIP-seq experiments using our fully enumerative GPU-based de novo motif discovery method, Argo_CUDA. Analysis revealed sets of significant IUPAC motifs corresponding to the binding sites of the target and partner transcription factors. For these ChIP-seq experiments, multiple regression models were constructed, demonstrating a significant dependence of the peak scores on the presence in the peak sequences of not only highly significant target motifs but also less significant motifs corresponding to the binding sites of the partner transcription factors. A significant correlation was shown between the presence of the target motifs FOXA2 and the partner motifs HNF4G, which found experimental confirmation in the scientific literature, demonstrating the important contribution of the partner transcription factors to the binding of the target transcription factor to DNA and, consequently, their important contribution to the peak score.

Keywords: chromatin immunoprecipitation with massively parallel sequencing; transcription factor binding sites; IUPAC motifs; co-binding of transcription factors; composite elements; multiple regression

Citation: Vishnevsky, O.V.; Bocharnikov, A.V.; Ignatieva, E.V. Peak Scores Significantly Depend on the Relationships between Contextual Signals in ChIP-Seq Peaks. *Int. J. Mol. Sci.* **2024**, *25*, 1011. https://doi.org/10.3390/ijms25021011

Academic Editor: Ming Chen

Received: 31 October 2023
Revised: 13 December 2023
Accepted: 9 January 2024
Published: 13 January 2024

1. Introduction

All functions during the whole life cycle of living beings are controlled at the genetic level through gene expression. Transcription is the first step in this complex multi-step process, which determines whether a gene is expressed at a given time. The transcription level of a particular gene depends on the cell type, tissue and organ and is regulated depending on the stage of cell differentiation and the stage of organism development [1,2]. Transcription is regulated by a large number of transcriptional regulatory proteins [3,4]. Here, a special role is played by transcription factors (TFs), as they are able to specifically bind to the regulatory regions of genes, determining the composition of multiprotein complexes that form the unique transcription machinery of each gene [5,6].

The development of ChIP-seq methods [7–10] using antibodies for transcription factors has led to the fast-paced accumulation of large bodies of data on the areas where transcription factor binding sites (TFBSs) localize, and thus substantially increased the amount of data on gene regulatory regions. The main advantage of any ChIP-seq method is that it allows the researcher to obtain information on the genome-wide localization of DNA-TF interaction sites in different tissues [11,12] at different stages of organism development [13–15] and under different external influences [16] in in vivo settings. For any target TF, the genome-wide ChIP-seq analysis normally reveals tens of thousands of

DNA sequences hundreds of nucleotides in length, corresponding to the genomic DNA fragments that have bound to the TF.

Raw data coming from ChIP-seq experiments are stored in data repositories such as GEO [17], ArrayExpress [18], ENA [19] and SRA [20] and are then fed to ChIP-seq analysis pipelines [7,21,22], utilizing peak callers [23], programs that map sequence reads onto a reference genome and identify the regions with the highest coverage on it; these regions being called the "ChIP-seq peaks". For each ChIP-seq peak, the peak callers make it possible to evaluate the peak score associated with the significance of this peak. The raw data accumulated in ChIP-seq data repositories served as the basis for second-level databases, which contain the results of ChIP-seq experiments classified according to species of organisms, tissue types and cell lines (CODEX [24], BloodChIP [25], hmChIP [26], CistromeDB [27], GTRD [28], ChIP-Atlas [29], TFBSbank [30] and Factorbook [31]), they also generate quality control metrics for the ChIP-seq experiments and have easy-to-use visual interfaces to analyze the localization of ChIP-seq peaks on genomic sequences.

Unfortunately, the peak sequences obtained from a ChIP-seq experiment are much longer than the binding site of any TF, which prevents the accurate localization of the binding site of the target TF and the determination of its size [21,32]. Two classes of computer methods are used to determine the exact locations of potential transcription factor binding sites in ChIP-seq peak sequences. One is based on the identification of ChIP-seq peak sequence regions which are significantly similar to position–weight matrices (PWMs) that describe the binding sites of known TFs and are stored in databases such as HOCOMOCO [33], JASPAR [34] and TRANSFAC [35]. To identify potential TFBSs using position–weight matrices, programs such as HOMER [36], MATCH [37] and MEME Suite [38] are used. Position–weight matrices do not normally consider interpositional dependencies [39,40]. It has been shown that considering dinucleotide dependencies [41], hidden Markov models (HMM) [42,43] and HMM-based TF flexible models (TFFMs) [44] improves the quality of TFBSs recognition in ChIP-seq peak sequences.

The other approach is based on the de novo detection of significantly overrepresented contextual signals in the sets of the DNA sequences in question. Once the contextual signals have been detected, the degree of their similarity with the position–weight matrices of known TFs stored in the corresponding databases is determined. The use of exhaustive de novo methods that guarantee finding the global optimum in the big data corresponding to the most significant DNA motif is time-consuming and expensive. This explains the researchers' inclination to consider only a small portion of randomly selected sequences [45] and to use heuristic approaches which can be roughly classified into four types: enumerative [46–50], probabilistic [51–56], nature-inspired [57,58] and deep learning [59]. In ChIP-seq peak sequences, such methods can very efficiently identify highly conserved and highly represented contextual signals corresponding to the binding sites of the target TF. At the same time, the efficient identification of more degenerate and poorly represented contextual signals corresponding to the binding sites of partner transcription factors among such data remains an unresolved problem.

Despite the existence of a large number of methods for identifying potential TFBSs, many authors note that the signal of the target factor is not detected in all ChIP-seq peaks [13,60–62]. This observation could be due to several factors. (1) TFBSs information stored in the databases is incomplete and the currently available computer-aided methods for identifying TFBSs in ChIP-seq peak sequences do not perform well enough. (2) The physical properties and structural features of DNA, such as melting enthalpy of DNA, DNA bending, groove width, etc., [63–68] are ignored. (3) Most computer-aided methods used for the identification of TFBSs are blind to the possibility that more than one variant of the binding site may exist for one transcription factor, which may either be indicative of the presence of more than one conformation of a DNA-protein complex, which, for example, is known for the transcription factors from SREBP family [69], or reflect the different binding preferences of a transcription factor due to post-translational modifications [70–73].

(4) In some cases, target factors may bind to DNA not directly, but indirectly, through protein–protein interactions with partner TFs and chromatin proteins [62].

ChIP-seq peak sequences are characterized by a high-level complexity of contextual organization and may either contain several potential binding sites for target TFs or not contain them at all. Additionally, they may contain quite a few potential binding sites for partner TFs. A wealth of experimental data provide evidence that transcription factors regulate transcription in close cooperation with each other and with other regulatory proteins (transcriptional co-activators and transcriptional co-repressors) [74,75]. This cooperation may result in the formation of, for example, the so-called enhanceosome, when several TFs interacting with a short sequence of a regulatory region form a common surface, which serves as a signal for recruiting coregulatory proteins [76,77]. In the analysis of regulatory regions, this situation corresponds to the presence of stable (frequent) combinations of closely spaced binding sites for transcription factors called the "*cis*-regulatory modules" (CRMs) [78] or composite elements (CEs) [79,80]. Examples of pairs of functionally interacting sites were originally presented in the COMPEL database [79,81]. Interactions between TFs binding to such pairs of sites may give rise to some subtle features of tissue- and stage-specific gene expression [82,83]. It has been shown that transcription factors can bind to the regulatory regions of genes in synergy [84,85] or in competition [86]. Some TFs can facilitate DNA binding to other TFs both by remodeling the nucleosomes and freeing up the regulatory region of the gene for binding to these TFs (as do pioneer TFs, such as FOXA2, Sp1, and PU.1 [87–89]), and through direct protein–protein interactions with these TFs [90,91].

Thus, taking into account pairwise and groupwise interactions between transcription factors is necessary for a deeper understanding of gene regulation and for scoring the observed ChIP-seq peaks. Such potentially interdependent TFs are normally sought when using computer-aided methods for the identification of significantly co-occurring contextual motifs corresponding to TFBSs. One group of works addressing this problem relies on the detection of all possible potential TFBSs in the ChIP-seq peak sequences of a particular transcription factor using PWMs from the JASPAR [34], HOCOMOCO [33] and TRANSFAC [35] databases [92–95]. The other group of works relies on the computer-aided analysis of the whole-genome distribution of ChIP-seq peaks for two TFs from different experiments on the assessment of non-randomness in the intersection of the peaks of these TFs, and the subsequent finding of potential TFBSs in the regions of the peak intersections on the basis of PWMs [96–98]. The main limitation of both approaches is the use of a limited set of PWMs for the recognition of partner TFBSs in ChIP-seq peaks; in addition, the second approach requires two independent ChIP-seq experiments under the same cellular conditions. De novo motif discovery methods make it possible to identify all significant contextual signals corresponding to the binding sites in ChIP-seq peak sequences, and target and partner TFs without the need to use PWM databases.

Despite a large number of methods for identifying TFBSs [36–38] and peak callers [23] for detecting ChIP-seq peaks and evaluating their peak scores, no computer-aided methods for estimating the dependence of the peak score of ChIP-seq peaks on the presence of various contextual signals in them are known.

In this work, we have analyzed the contextual organization of ChIP-seq peak sequences in experiments with 10 TFs belonging to six different superclasses corresponding to the types of their DNA-binding domains [99]. The analysis was carried out using a modified version of our original software package Argo_CUDA version 2.0 [100], with which it is possible to identify sets of significant contextual motifs in the samples of DNA sequences written in a 15-letter IUPAC code [101]. Analysis of the data obtained from each of the ChIP-seq experiments revealed the sets of significant IUPAC motifs (target motifs) corresponded to the binding sites of the target TF studied in each particular experiment and significant IUPAC motifs corresponding to the binding sites of other partner TFs (partner motifs). The motifs recognized in all the ChIP-seq experiments formed the basis of multiple regression

models that demonstrated the significant dependence of peak scores of the ChIP-seq peak sequences on the presence of IUPAC motifs in these sequences.

2. Results

2.1. Preparing Sets of ChIP-Seq Peak Sequences

In this work, we have analyzed the ChIP-seq peak sequences obtained by different research groups [11–15,102–106] in experiments conducted to locate the binding regions for 10 transcription factors (Table 1) in genomic DNA from the CistromeDB database [27]. All ChIP-seq data analyzed in this work were obtained on mouse cells or cell lines. The eighth column of Table 1 contains information about the cell line or type. ChIP-seq experiments were carried out on liver cells [11,102,104–106], bone marrow-derived macrophages [12], embryonic stem cells [13,14], erythroid cells from fetal livers [103] and differentiating myoblasts C2C12 [15]. For the purpose of analysis, only experiments with the highest quality scores and containing at least 5000 ChIP-Seq peak sequences were considered. We considered transcription factors that had PWM models presented in the HOCOMOCO database [33] and belong to different TF superclasses according to Wingender's classification [99]. Only transcription factors from six TF superclasses met these conditions.

Table 1. Brief description of ChIP-seq data analyzed in this work.

TF [1]	ID [2]	N [3]	N_L [4]	N_C [5]	TF Superclass [6]	TF Family [7]	Cells Type [8]	Ref. [9]
CEBPA	39908	33,559	5000	28,559	1. Basic domains	1.1.8. CEBP-related	Mouse liver cells	[11]
CEBPB	72841	13,374	5000	8374	1. Basic domains	1.1.8. CEBP-related	Mouse liver cells	[102]
NFE2L2	70563	27,065	5000	22,065	1. Basic domains	1.1.1. Jun-related	Mouse bone marrow-derived macrophages	[12]
SP1	47755	24,404	5000	19,404	2. Zinc-coordinating DNA-binding domains	2.3.1. Three-zinc finger Krüppel-related	Mouse embryonic stem cells	[13]
GATA1	46419	7534	5000	2534	2. Zinc-coordinating DNA-binding domains	2.2.1. C4-GATA-related	Mouse erythroid cells of fetal liver	[103]
FOXA2	3266	25,191	5000	20,191	3. Helix-turn-helix domains	3.3.1. FOX	Mouse liver cells	[104]
FOXO1	92461	11,433	5000	6433	3. Helix-turn-helix domains	3.3.1. FOX	Mouse liver cells	[105]
NFYA	48618	5975	5000	975	4. Other all-alpha-helical DNA-binding domains	4.2.1. Heteromeric CCAAT-binding	Mouse embryonic stem cells	[14]
MEF2D	38097	34,789	5000	29,789	5. Alpha-Helices exposed by beta-structures	5.1.1. Regulators of differentiation	Mouse differentiating myoblasts C2C12	[15]
STAT5B	5839	18,510	5000	13,510	6. Immunoglobulin fold	6.2.1. STAT	Mouse liver cells	[106]

[1] Name of the transcription factor; [2] ID of ChIP-seq experiment in the CistromeDB database [27]; [3] Total number of peaks identified in the ChIP-seq experiment and checked for the absence of non-canonical symbols; [4] Number of ChIP-seq peak DNA sequences included in the training set; [5] Number of ChIP-seq peak DNA sequences included in the control set; [6] Number and name of the superclass to which the TF belongs according to Wingender's classification [99]; [7] Number and name of the family to which the TF belongs according to Wingender's classification [99]; [8] Organism and cell type used in the ChIP-seq experiment; [9] Link to relevant publication.

The transcription factors we have chosen regulate the vital functions of cells, including growth, division and differentiation. The transcription factors CEBPA and CEBPB from the C/EBP family are involved in the regulation of the cell cycle and differentiation of various cell types, including blood, liver and adipose tissue cells [107]. The transcription factor NFE2L2/NRF2 is activated in response to inflammation and cell damage, and regulates the expression of antioxidant defense genes [108]. NFE2L2 KO mouse experiments showed

that this TF controls the development of the small intestine [109]. SP1 is a ubiquitously expressed transcription factor involved in the control of erythroid cell specification [13]. GATA1 is known to regulate erythropoiesis [110]. FOXA2 is known to be a transcription factor with pioneering functions and is involved in the regulation of morphogenesis [111]. FOXO1, known to be a regulator of the cell cycle, apoptosis and oxidative stress, is involved in the regulation of placental and cardiovascular morphogenesis [112,113]. NFYA controls the expression of multiple genes involved in cell cycle regulation and steps down as a regulator of the stemness and proliferation of mouse embryonic stem cells (mESCs) and human hematopoietic stem cells (hHSCs) [114]. MEF2D is involved in the differentiation of muscle cells [15]. STAT5B is involved in the regulation of the differentiation of osteoblasts, adipocytes and neuronal cells [115–117].

Based on the results of each ChIP-seq experiment, DNA sequences were obtained in the [−100;+100] region relative to the maxima of the ChIP-seq peaks. Next, the ChIP-seq peaks were ranked according to the peak scores (PSs), which reflect the enrichment of the peaks. Larger numbers of peak scores represent more confident peak calls.

To ensure that the comparative characteristics of IUPAC motifs recognized do not depend on the size of the training sets, we selected 5000 peak sequences considering the maximum peak scores for each of the 10 ChIP-seq experiments. From these sequences, training sets were compiled. The remaining sequences were included in the control sets for each ChIP-seq experiment. DNA sequences corresponding to ChIP-seq peaks were extracted from the GRCm38_97 mouse genomic assembly available in the EMBL database according to CistromeDB annotation [118].

As can be seen from Table 1, the numbers of entries in the control sets of ChIP-seq peak sequences ranged from 975 for NFYA to 29,789 for MEF2D.

2.2. Identification of Significant Oligonucleotide Motifs in the ChIP-Seq Sequences

In the training data sets created for each of the ChIP-seq experiments and consisting of 5000 DNA sequences located in the [−100;+100] regions relative to the ChIP-seq peaks, we de novo identified sets of significant degenerate motifs written in the 15-letter IUPAC code (A,T,G,C, R = G/A, Y = T/C, M = A/C, K = G/T, W = A /T, S = G/C, B = T/G/C, V = A/G/C, H = A/T/C, D = A/T/G, N = A/T/G/C). The identification of significant IUPAC motifs was carried out using our original de novo motif discovery system Argo_CUDA [100] with some improvements (see Section 4). Unlike heuristic methods, the de novo motif discovery GPU-based Argo_CUDA assesses the significance of all possible IUPAC motifs of a given length, which guarantees that a global optimum will be found. At the same time, IUPAC motifs were considered to be significant if they were significantly overrepresented in the set of sequences being analyzed compared to the values of abundance expected to have been observed for random reasons. A detailed description of the significance criterion (1) and the boundary values used are also provided in the Section 4.

As a result of the analysis, sets of significant IUPAC motifs were identified in each of the ten training sets corresponding to ChIP-seq experiments with different transcription factors (see Table 2).

Table 2 shows that although the number of DNA sequences in the sets of the results of different ChIP-seq experiments was the same (N = 5000), and the same boundary values were used for the significance criterion (1), the number of motifs identified in each ChIP-seq experiment and their maximum significance vary strongly. Thus, there was an ~1.8-fold variation in the number of all motifs identified (at $P_{Bonf}(n,N) < p_0 = 10^{-2}$) in these ChIP-seq experiments. The largest number of significant motifs (270) was found for NFYA, and the least one (151) was found for CEBPA. An average of 206 IUPAC motifs were identified in all 10 ChIP-seq data sets. The highest maximum significance of the motifs was observed in the ChIP-seq experiment with FOXA2 ($P_{Bonf}(n,N) = 10^{-1462}$), and the lowest, with FOXO1 ($P_{Bonf}(n,N) = 10^{-447}$). As can be noted, fewer significant motifs were in the training set **FOXA2$_{\mathbf{L}}$** (202) than in the training set **FOXO1$_{\mathbf{L}}$** (220). On the whole, there was a negative correlation (r = −0.39) between the number of motifs detected and their maximum significance. For

each of the sets, we inferred the number of the most significant (P_{Bonf} $(n,N) < p_0 = 10^{-30}$) motifs detected and its ~2.6-fold variation. The maximal number of the most significant motifs (81) was found for NFYA, and the minimum number (31) was found for for GATA1. An average of 54 most-significant IUPAC motifs were detected in all 10 ChIP-seq datasets. As can be noted, the number of the most significant motifs detected in each ChIP-seq experiment strongly correlates with the total number of all motifs detected in it (r = 0.95) and negatively correlates with the maximum significance of the motifs detected (r = −0.31).

Table 2. Characteristics of significant IUPAC motifs identified in the ChIP-seq peak DNA sequences for ten target transcription factors. Information is provided both on all motifs detected in each IUPAC motif set ($p_0 = 10^{-2}$) and on the most significant of them ($p_0 = 10^{-30}$).

Transcription Factor	Maximum Significance of the Motifs, $-log_{10}(P_{Bonf}\ (n,N))$	Number of Motifs Identified at $P_{Bonf}\ (n,N) < p_0 = 10^{-2}$	Number of Motifs Identified at $P_{Bonf}\ (n,N) < p_0 = 10^{-30}$
CEBPA	1585	151	36
CEBPB	1028	164	39
NFE2L2	880	239	71
SP1	520	243	80
GATA1	932	168	31
FOXA2	1462	202	50
FOXO1	447	220	51
NFYA	1134	270	81
MEF2D	749	219	57
STAT5B	744	181	44

The dependence of the number of IUPAC motifs detected in each of the sequence sets on the motif significance is shown in Figure 1.

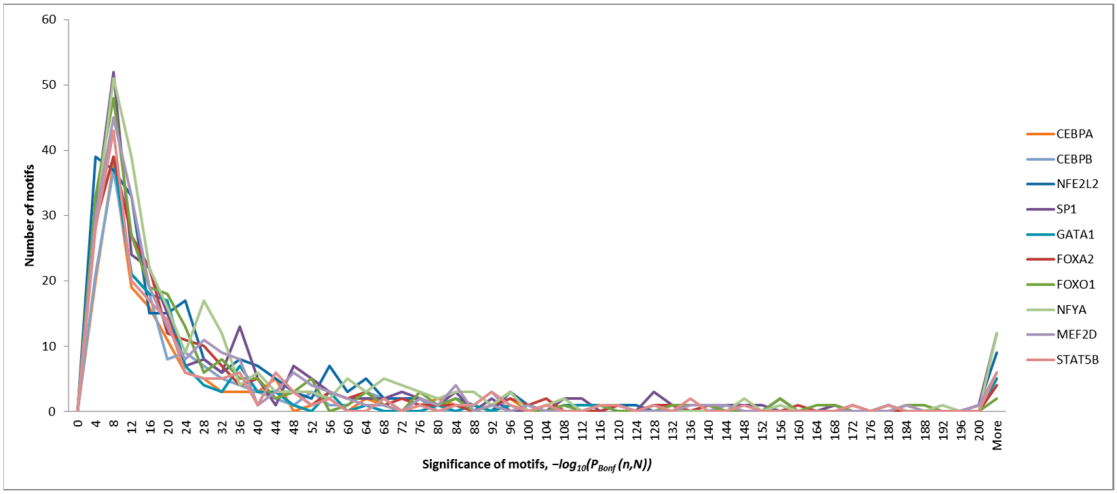


Figure 1. Dependence of the number of IUPAC motifs (*y*-axis) detected in ChIP-seq experiments with 10 transcription factors on the significance of these motifs (*x*-axis).

As can be noted, all the sets of motifs we obtained had a similar distribution pattern: the region corresponding to the probability of observing them for random reasons P_{Bonf} (n,N) ϵ $[10^{-30}; 10^{-2}]$ has a pronounced peak in the number of the motifs detected, which further turns into a very long tail.

Table 3 shows examples and characteristics of the most significant (P_{Bonf} $(n,N) < 10^{-30}$) IUPAC motifs detected in the training set **FOXA2$_L$** for ChIP-seq peak sequences in the FOXA2 experiment (ID 3266 in CistromeDB). FOXA2, known as hepatocyte nuclear factor 3-beta (HNF-3B), belongs to the FOXA subfamily. This subfamily, in turn, belongs to the FOX family, whose proteins contain a relatively conserved DNA binding domain known as the winged-helix or forkhead domain.

Table 3. Examples and characteristics of the most significant (P_{Bonf} $(n,N) < 10^{-30}$) IUPAC motifs detected with Argo_CUDA in the training set of ChIP-seq peak sequences in the experiment involving the transcription factor FOXA2.

Motif	Actual Abundance [1], F	Expected Abundance [2], Q	Significance [3], $-log_{10}(P_{Bonf}\ (n,N))$
TRTWKACH	0.67	0.15	1462
RTTKACHY	0.54	0.20	620
TMAAYANS	0.54	0.26	395
TWKACHYW	0.55	0.27	367
TTKRTYTW	0.30	0.14	172
TKAHYTWK	0.46	0.27	157
TRTTKRTY	0.30	0.15	139
RAAYHAAY	0.31	0.17	128
TTRNGHAA	0.29	0.16	116
KACDTWGN	0.30	0.17	103
TAAHYABW	0.32	0.19	91
ARMYAAGV	0.30	0.18	81
TGYGTACH	0.09	0.03	77
TRTWTGCW	0.15	0.07	69
AAAMAAAR	0.13	0.06	60
ATMMAYAN	0.28	0.18	55
WTRTTTGY	0.19	0.11	54
TCRAYADW	0.19	0.11	52
GTACRCAH	0.06	0.02	46
TTYGCTYW	0.12	0.06	45
TNGCTHWG	0.24	0.16	40
TNACYMWG	0.30	0.22	31

[1] F is the proportion of sequences observed to contain at least one IUPAC motif; [2] Q is the proportion of sequences expected to contain the IUPAC motif for random reasons; [3] $-log_{10}(P_{Bonf}\ (n,N))$ is the Bonferroni-corrected binomial probability of observing motifs for random reasons (see Section 4).

For example, the most significant motif TRTWKACH = (T)(A/G)(T)(A/T)(G/T)(A)(C)(A/T/C) detected in **FOXA2$_L$** was present in n = 3345 out of N = 5000 sequences (F ~ 0.67), the expectation to observe it for random reasons was 764 sequences (Q ~ 0.15), and the Bonferroni-corrected binomial probability of observing it in at least 3345 out of 5000 sequences for random reasons was $P_{Bonf}\,(3345, 5000) = 10^{-1462.3}$.

2.3. Multiple Regression Model-Based Assessment of the Dependence of ChIP-Seq Peak Scores on the Presence of Significant IUPAC Motifs in Them

To assess the relationship between the presence of motifs in a ChIP-seq peak sequence and the natural logarithm of their peak scores (*ln(PS)*), multiple regression models were built (see Section 4) on the training sets for each ChIP-seq experiment using STATISTICA (StatSoft^{TM}, Tulsa, OK, USA). The regression coefficients obtained were then used to assess the correlation between the peak scores predicted by the multiple regression models and the peak scores that were actually observed in the training and control sets for each ChIP-seq experiment. The results of the analysis and the correlation coefficients obtained for the training and control sets are shown in Table 4.

Table 4. The results of assessing the dependence of the natural logarithm of peak scores (*ln(PS)*) on the presence of significant ($p < 10^{-2}$) IUPAC motifs in the DNA sequences of the training and control sets for ChIP-seq experiments with 10 transcription factors. The multiple regression coefficients for the training sets were calculated using STATISTICA (StatSoft^{TM}, Tulsa, OK, USA).

		Training Set			Control Set		
Transcription Factor	**Number of Significant Motifs**	**Number of Sequences in the Set, N**	**Correlation Coefficient, r**	**Max *ln(PS)***	**Number of Sequences in the Set, N**	**Correlation Coefficient, r**	**Max *ln(PS)***
CEBPA	151	5000	0.38	5.29	28,559	0.11	3.6
CEBPB	164	5000	0.37	5.53	8374	0.13	2.74
NFE2L2	239	5000	0.33	5.54	22,065	0.01 *	3.17
SP1	243	5000	0.57	5.69	19,404	0.06	3.5
GATA1	168	5000	0.3	5.45	2534	−0.03 *	2.42
FOXA2	202	5000	0.33	5.4	20,191	0.09	3.49
FOXO1	220	5000	0.32	5.68	6433	0.08	2.45
NFYA	270	5000	0.53	5.46	975	0.07	1.88
MEF2D	219	5000	0.36	5.41	29,789	0.07	2.93
STAT5B	181	5000	0.32	5.2	13,510	0.09	3.1

* Non-significant ($p > 0.05$) correlation coefficients r in assessing the relationship between the predicted and observed values of peak scores, with the Bonferroni correction [119] taken into account.

2.3.1. Assessment of the Dependence of Peak Scores on the Presence of Significant IUPAC Motifs in ChIP-Seq Peaks in the Experiment with the Transcription Factor FOXA2

Let us consider in detail the construction of a multiple regression model, using the analysis of ChIP-seq data from the FOXA2 experiment as an example. For each of the 5000 DNA sequences of the ChIP-seq peaks in the [−100;100] region relative to its maximum value in the training set **FOXA2$_L$**, the abundance of each of the 202 motifs previously detected in this IUPAC motif set was assessed. The abundance of a motif in a sequence means the number of observations of the given motif along the entire length of the DNA sequence in both of its strands. The abundance vectors, obtained in this way for all motifs considered, served as independent variables. The dependent variable was information about the natural logarithm of the peak scores of the sequences in the training set. To assess the dependence of peak scores on the abundance of IUPAC motifs in these ChIP-seq peaks, a multiple regression model was built using STATISTICA (StatSoft^{TM}, Tulsa, OK, USA). Table 5 shows examples of independent variables with significant ($p < 0.05$) regression coefficients and their characteristics. A complete table of the regression coefficients for all IUPAC motifs considered is provided in "Supplementary Materials", Table S1.

Table 5. Examples of independent variables with significant ($p < 0.05$) regression coefficients obtained when constructing a multiple regression model to assess the dependence of the *ln*(*PS*) value in the ChIP-seq experiment with the transcription factor FOXA2 on the presence of IUPAC motifs in these peaks. Multiple regression coefficients were calculated using STATISTICA (StatSoft^{TM}, Tulsa, OK, USA).

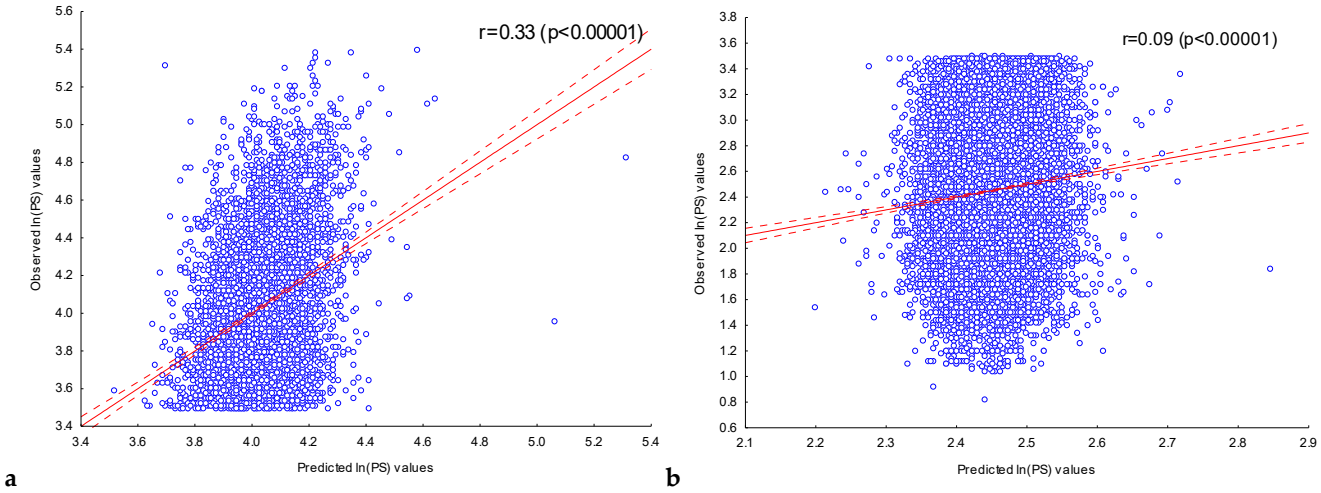
Independent Variable	**Regression Coefficient**	***p*-Value**
TRTWKACH	0.064945	0.000003
RTTKACHY	0.027893	0.039629
TMAAYANS	0.057817	0.0000001
TWKACHYW	0.027454	0.006242
TTRNGHAA	0.024021	0.031108
KACDTWGN	0.028459	0.011256
TAAHYABW	−0.027147	0.006997
CCRCCCCB	−0.081117	0.000936
CGBTCGVN	0.118823	0.006146
TTSGYWRN	0.018182	0.040455

Table 5. *Cont.*

Independent Variable	Regression Coefficient	*p*-Value
AACAWGVV	0.032637	0.004226
TRATTRRY	0.037430	0.032885
TTNRTTCW	−0.035167	0.016293
GGWGRVHG	0.024464	0.008203
WSCSTRKS	0.018255	0.033926
HCGBTCGV	−0.113689	0.043413
GGCRGGAV	0.046776	0.009803
TTKACWRA	0.033587	0.035103
GGHNGAGH	0.021662	0.040533
KRAGCBAN	0.026475	0.009812
ACVCWRMS	0.023771	0.010771
WCCCCVVC	0.04015	0.01686
AMVCAYAG	0.03683	0.00944
CGNMYCGG	0.08602	0.008174
CKTCCGKN	0.0739	0.044389
CCTSGRMK	0.036704	0.015545
TGTGGACW	0.103851	0.000909
GSARHGGR	−0.023597	0.040487
WGCGGYSG	0.084659	0.021616
TWWKTAAY	−0.053886	0.001453

Of the total number of independent variables considered that corresponded to the abundance of 202 significant IUPAC motifs, only 30 (15%) had significant ($p < 0.05$) regression coefficients.

Figure 2a shows the dependence of the observed value of $ln(PS)_{Obs}$ for the ChIP-seq peak sequences in the training set **FOXA2$_{\mathbf{L}}$** on the expected value of their $ln(PS)_{Exp}$ calculated with the multiple regression model using all 202 independent variables. The observed $ln(PS)_{Obs}$ and expected $ln(PS)_{Exp}$ values were significantly ($p < 10^{-5}$) correlated (r = 0.33).



a **b**

Figure 2. Dependence of the observed value of $ln(PS)_{Obs}$ for ChIP-seq peak sequences in the training (**a**) and control (**b**) FOXA2 sets on the expected value. The expected value of $ln(PS)_{Exp}$ for the peaks was predicted using the multiple regression model on the presence of 202 significant IUPAC motifs previously detected in the training set **FOXA2$_{\mathbf{L}}$**. Solid and dashed lines represent the regression line and the bounds of its 95% confidence interval as calculated using STATISTICA (StatSoft^{TM}, Tulsa, OK, USA). r is the linear correlation coefficient and p is its statistical significance.

This regression model was used to predict the expected $ln(PS)_{Exp}$ values for ChIP-seq peak sequences in the control set **FOXA2$_{\mathbf{C}}$**. Figure 2b shows the plotted relationship

between the observed $ln(PS)_{Obs}$ and the expected $ln(PS)_{Exp}$ values predicted. For these values, a significant ($p < 10^{-5}$) correlation was also observed (r = 0.09).

Thus, the information about the presence of ChIP-seq peaks in the DNA sequences in the FOXA2 experiment allows conclusions to be drawn about the peak score values for the ChIP-seq peaks of FOXA2, which had not previously been used for detecting significant IUPAC motifs or for constructing multiple regression models.

2.3.2. Assessment of the Dependence of Peak Scores in Experiments with Nine Transcription Factors on the Presence of Significant IUPAC Motifs in Their ChIP-Seq Peaks

Similarly, multiple regression models were built on the training sets of the remaining nine transcription factors. These models were used to predict the peak score of ChIP-seq peak sequences in the training and control sets (Table 4, "Supplementary Materials", Figure S1). The analysis performed showed that the correlation coefficients r in predicting the observed values of peak scores were significant for all ten training sets. They ranged from r = 0.3 for the **GATA1$_L$** set to r = 0.57 for the **SP1$_L$** set. On average, the training sets were characterized by r = 0.38. The r value for each of the training sets correlated non-significantly ($p > 0.05$) with the number of previously detected significant IUPAC motifs in it (r = 0.58).

The vast majority (8 out of 10) of the control sets had significant correlation coefficients in predicting the observed values of peak scores. At the same time, it was not possible to build a reliable prediction model for the **NFE2L2$_C$** and **GATA1$_C$** sets. On average, significant correlation coefficients were r = 0.09 and ranged from r = 0.06 for the **SP1$_C$** set to r = 0.13 for the **CEBPB$_C$** set. A comparison of the correlation coefficients r for the training and control sets shows that the values were significantly lower for the latter. The analysis performed did not reveal significant ($p < 0.05$) dependencies of the r values for the control sets on the size of these sets or on the r values for the corresponding training sets or on the maximum peak score in the sets.

Thus, it can be hypothesized that the DNA sequences in the ChIP-seq peaks with the highest peak scores are consistent with a model of context organization which is different to that for ChIP-seq peaks with low peak scores.

3. Discussion

3.1. Assessment of the Contribution of the Most Significant IUPAC Motifs to the Prediction of the Peak Scores

As noted previously in the analysis of Figure 1, all of the significant IUPAC motifs detected in each training set were divided into two groups: (1) the largest group, consisting of IUPAC motifs whose probability of being observed was $P_{Bonf}(n,N) \in [10^{-30}; 10^{-2}]$ and (2) the most significant motifs ($P_{Bonf}(n,N) < 10^{-30}$) in the tail of the distribution. It can be assumed that taking the presence of only the most significant IUPAC motifs in the DNA sequence of a ChIP-seq peak into account will be sufficient to predict its peak score. In order to assess how strongly the peak score of a ChIP-seq peak sequence depends on the presence of only the most significant IUPAC motifs in it, we built a multiple regression model for predicting the peak score based on IUPAC motifs only from the group with $P_{Bonf}(n,N) < 10^{-30}$. The model was built for the sets with ChIP-seq peak sequences obtained in the experiments with FOXA2 and SP1.

3.1.1. Assessment of the Dependence of the Peak Scores in the Experiment with the Transcription Factor FOXA2 on the Presence of Only the Most Significant IUPAC Motifs in the Peaks

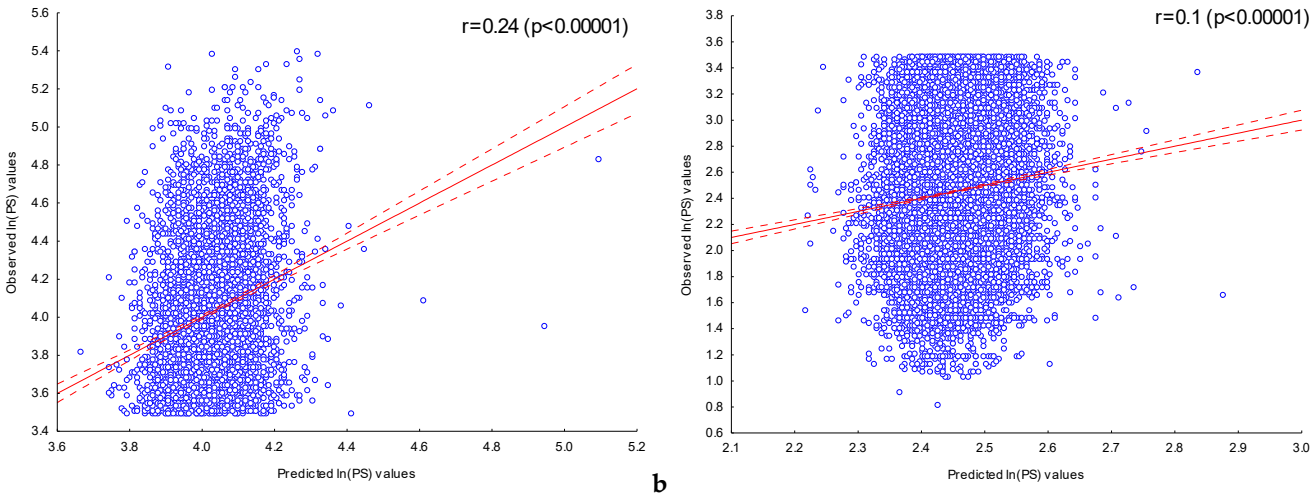
A multiple regression model was built for the **FOXA2$_L$** training set with STATISTICA (StatSoft^{TM}, Tulsa, OK, USA) using only the 50 most significant ($P_{Bonf}(n,N) < 10^{-30}$) IUPAC motifs detected in it (the motifs are listed in the "Supplementary Data"). Table 6 shows examples of independent variables with significant ($p < 0.05$) regression coefficients and their characteristics. A complete table of the regression coefficients for all IUPAC motifs considered is provided in the Supplementary Materials, Table S2.

Table 6. Examples of independent variables with significant ($p < 0.05$) regression coefficients obtained when constructing a multiple regression model to assess the dependence of the *ln(PS)* value in the ChIP-seq experiment with the transcription factor FOXA2 on the abundance of the most significant (P_{Bonf} $(n,N) < 10^{-30}$) IUPAC motifs. Multiple regression coefficients were calculated using STATISTICA (StatSoft^{TM}, Tulsa, OK, USA).

Independent Variable	Regression Coefficient	*p*-Value
TRTWKACH	0.066537	0.000001
TMAAYANS	0.051297	0.0000001
TWKACHYW	0.021614	0.028887
KACDTWGN	0.025535	0.015482
TAAHYABW	−0.033779	0.000422
ARMYAAGV	0.02232	0.020651
CCRCCCCB	−0.056879	0.017154
CGSCBBCG	0.040932	0.014188
GCGYKCGN	0.115472	0.009763
CGMRCGCV	−0.131678	0.009819
ANCRAHGV	0.018453	0.040805

As can be seen from Table 6, of all the independent variables corresponding to the abundance of the 50 most significant IUPAC motifs, only 11 (22%) had significant ($p < 0.05$) regression coefficients.

With the multiple regression model constructed, peak scores were predicted for the sequences in the training set **FOXA2$_L$** and the control set **FOXA2$_C$**. Figure 3 shows the dependencies of the observed $ln(PS)_{Obs}$ values for the ChIP-seq peak sequences of the training set **FOXA2$_L$** (Figure 3a) and the control set **FOXA2$_C$** (Figure 3b) on their expected $ln(PS)_{Exp}$ values calculated with the multiple regression model constructed, using all 50 independent variables.



a **b**

Figure 3. Dependence of the observed values of $ln(PS)_{Obs}$ for ChIP-seq peak sequences from (**a**) the training set **FOXA2$_L$** and (**b**) the control set **FOXA2$_C$** on their expected values. The expected value of $ln(PS)_{Exp}$ for the ChIP-seq peaks was predicted using a multiple regression model for the presence of the most significant (P_{Bonf} $(n,N) < 10^{-30}$) IUPAC motifs previously detected in the training set **FOXA2$_L$**. Solid and dashed lines represent the regression line and the bounds of its 95% confidence interval as calculated using STATISTICA (StatSoft^{TM}, Tulsa, OK, USA). r is the linear correlation coefficient and *p* is its statistical significance.

For both sets, the observed $ln(PS)_{Obs}$ and expected $ln(PS)_{Exp}$ values were significantly ($p < 10^{-5}$) correlated. At the same time, for the training set **FOXA2$_L$**, the use of only the most significant (P_{Bonf} $(n,N) < 10^{-30}$) IUPAC motifs led to a lower correlation (r = 0.24)

than the use of all 202 significant motifs (P_{Bonf} (n,N) < 10^{-2}), r = 0.33. On the other hand, taking into account the localization of only the most significant (P_{Bonf} (n,N) < 10^{-30}) IUPAC motifs for the control set **FOXA2$_C$** allowed us to obtain a slightly higher correlation (r = 0.1) than was obtained using all 202 significant IUPAC motifs (r = 0.09). On the whole, we can conclude that using information about localization in the ChIP-seq sequences of only the most significant motifs without taking into account other motifs does not allow predicting the peak score values with the greatest accuracy in all cases.

3.1.2. Assessment of the Dependence of Peak Scores for Peak Sequences in the Experiment with the Transcription Factor SP1 on the Presence of Only the Most Significant IUPAC Motifs in Them

A multiple regression model was built for the training set **SP1$_L$** with STATISTICA (StatSoft^{TM}, Tulsa, OK, USA) using only the 80 most significant (P_{Bonf} (n,N) < 10^{-30}) IUPAC motifs detected in it (the motifs are listed in the "Supplementary Data"). Table 7 shows examples of independent variables with significant (p < 0.05) regression coefficients and their characteristics. A complete table of the regression coefficients for all IUPAC motifs considered is provided in the Supplementary Materials, Table S3.

Table 7. Examples of independent variables with significant (p < 0.05) regression coefficients obtained when constructing a multiple regression model to assess the dependence of the ln(PS) value in the ChIP-seq experiment with the transcription factor SP1 on the abundance of the most significant (P_{Bonf} (n,N) < 10^{-30}) IUPAC motifs. Multiple regression coefficients were calculated using STATISTICA (StatSoft^{TM}, Tulsa, OK, USA).

Independent Variable	Regression Coefficient	p-Value
ATTSGHYR	0.06863	0.009941
RRCSAATS	−0.074581	0.021391
TGTAGTYY	0.25583	0.0000001
CSAATSRV	0.141722	0.0000001
TACAWNTC	−0.103949	0.031145
ANWTGTAG	0.155459	0.000388
TBYBATTG	0.05046	0.00207
GGGANWTG	0.147136	0.0000001
WGGGYGGG	0.042234	0.038834
TKCYGGGW	0.051412	0.00338
CTTCCKGB	−0.046938	0.011891
RGGCGGGH	0.050394	0.001738
ATTSGYYY	0.076936	0.00512
TATTGGHY	0.172836	0.000022
GGGHSGWG	−0.025723	0.047439
CGGKRCBD	0.029378	0.019498
RABBGACR	0.075836	0.0000001
TTGGTCNR	0.079731	0.000773
TAGTYYWH	0.050178	0.015284
TTTRHWTW	−0.037338	0.037889
WKCAAAKN	−0.043257	0.004714
GTCAYGTG	−0.098339	0.001426
TGANTGAC	0.134395	0.000061
AVHGAYAR	−0.043841	0.001981
AYGATTSG	0.242444	0.0000001
GGATTSGH	0.121602	0.000005
ACGSAHGY	0.053019	0.026689
GVATKCTG	0.058493	0.038876
AGATAAGV	−0.129716	0.000021

As can be seen from Table 7, of the 80 independent variables considered, 29 had significant ($p < 0.05$) regression coefficients (36%).

With the multiple regression model constructed, peak scores were predicted for the sequences of the training set **SP1$_L$** and the control set **SP1$_C$**. Figure 4 shows the dependencies of the observed $ln(PS)_{Obs}$ values for the ChIP-seq peak sequences of the training set **SP1$_L$** (Figure 4a) and the control set **SP1$_C$** (Figure 4b) on their expected $ln(PS)_{Exp}$ value calculated with the multiple regression model constructed, using all independent variables corresponding to the abundances of the 80 most significant IUPAC motifs.

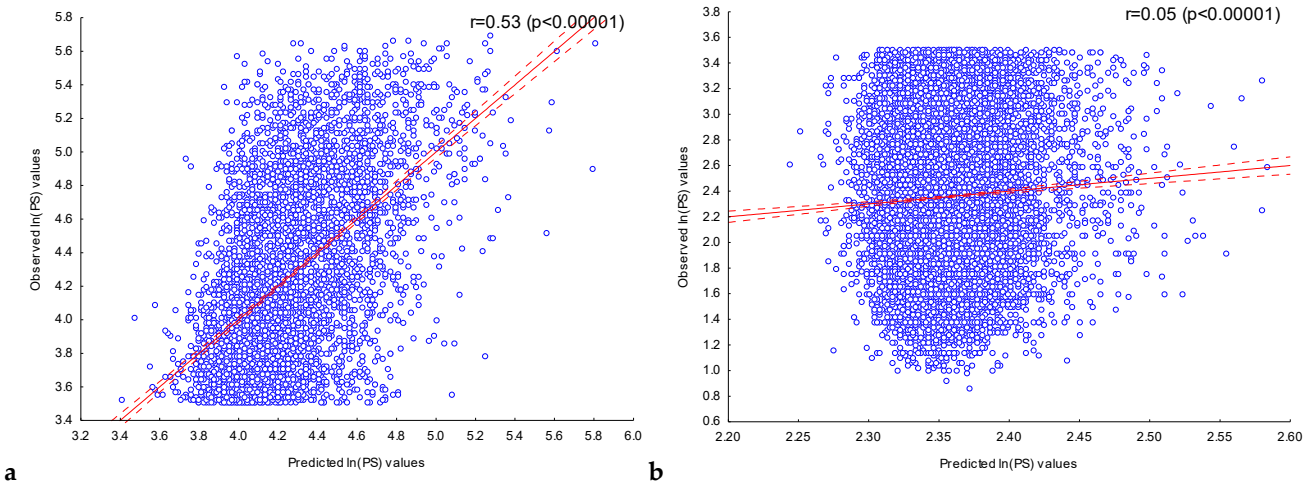


Figure 4. Dependence of the observed values of $\ln(PS)_{Obs}$ for ChIP-seq peak sequences from (**a**) the training set **SP1$_L$** and (**b**) the control set **SP1$_C$** on their expected values. The expected value of $\ln(PS)_{Exp}$ for the ChIP-seq peaks was predicted using a multiple regression model for the presence of the most significant ($P_{Bonf}(n,N) < 10^{-30}$) IUPAC motifs previously detected in the training set **SP1$_L$**. Solid and dashed lines represent the regression line and the bounds of its 95% confidence interval as calculated using STATISTICA (StatSoft^{TM}, Tulsa, OK, USA). r is the linear correlation coefficient and p is its statistical significance.

For both sets, the observed $ln(PS)_{Obs}$ and expected $ln(PS)_{Exp}$ values were significantly ($p < 10^{-5}$) correlated. In the case of the training set **SP1$_L$**, the use of only the 80 most significant ($P_{Bonf}(n,N) < 10^{-30}$) IUPAC motifs led to a somewhat lower correlation (r = 0.53) than the use of all 243 significant IUPAC motifs ($P_{Bonf}(n,N) < 10^{-2}$), r = 0.57 (Figure S1, Supplementary Materials, Table 4). In the case of the control set **SP1$_C$**, the use of information about the localization of only the 80 most significant ($P_{Bonf}(n,N) < 10^{-30}$) IUPAC motifs also led to a somewhat lower correlation (r = 0.05) than the use of all 243 significant IUPAC motifs ($P_{Bonf}(n,N) < 10^{-2}$), r = 0.06 (Figure S1, Supplementary Materials, Table 4).

Thus, the analysis of data on the ChIP-seq experiments with two TFs allows us to conclude that, to predict the peak scores of ChIP-seq peak sequences, it is necessary to use information about the localization of both the most significant motifs and other IUPAC motifs in them.

3.2. Assessment of the Correlations between the Presence of the Target IUPAC Motifs and Partner IUPAC Motifs in the Peak Sequences in the Experiment with Transcription Factor FOXA2

While we were building multiple regression models for all of the transcription factors considered, the analysis of the regression coefficients (Tables 5–7) suggested that only a small fraction of them were significant ($p < 0.05$). One of the reasons for this may be the internal interdependencies of the presence of different IUPAC motifs in the ChIP-seq peak sequences.

3.2.1. Assessment of the Correlations between the Co-Occurrences of the Most Significant (P_{Bonf} $(n,N) < 10^{-30}$) IUPAC Motifs in the Peak Sequences

To test the proposed assumption, we assessed the correlations between the co-occurrences of the most significant ($p < 10^{-30}$) motifs in the ChIP-seq peak sequences from the **FOXA2$_L$** set using the phi-coefficient [120] to assess paired correlations. Each IUPAC motif from the pair of motifs considered was assumed to be located in the ChIP-seq peak sequence if it had been found at least once in either of the DNA strands. Figure 5 shows a heat map of the pairwise correlations obtained (built with the Heatmapper tool [121] (see Section 4)).

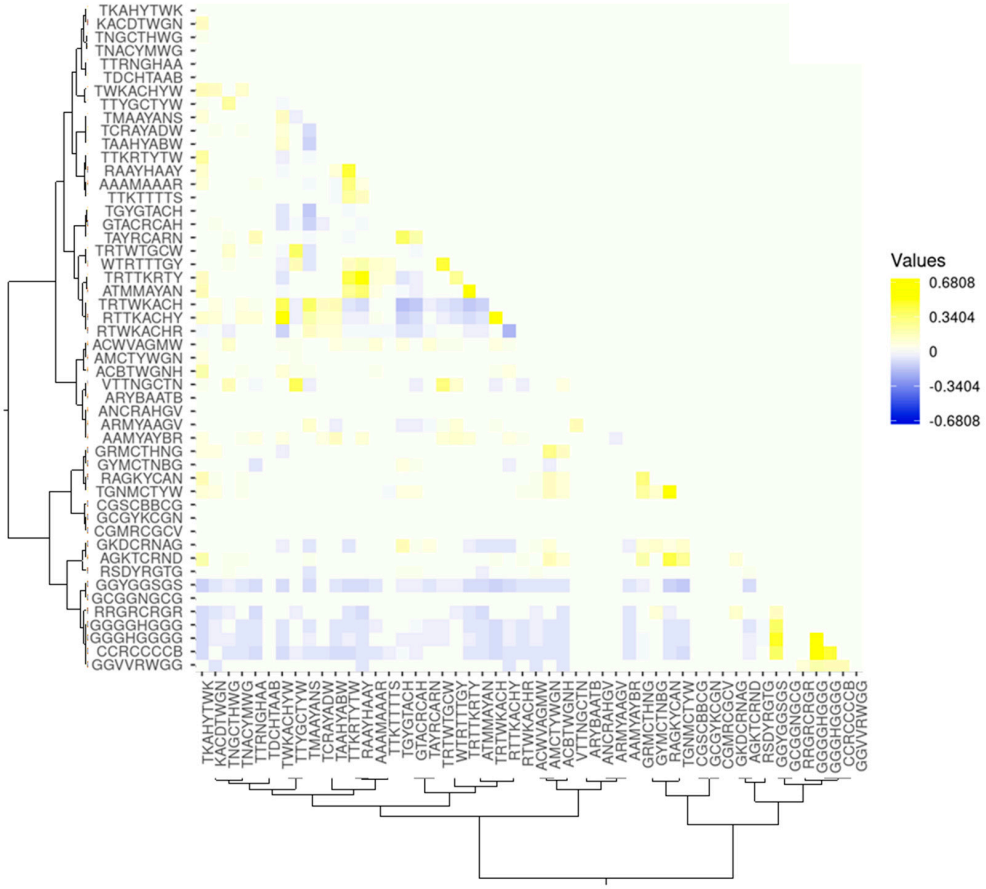


Figure 5. Heat map of significant ($p < 0.05$) pairwise correlations between the co-occurrence of the 50 most significant ($p < 10^{-30}$) IUPAC motifs found in [−100;+100] DNA sequences relative to the ChIP-seq FOXA2 peak maximum. Blue, a negative correlation; yellow, a positive correlation; white, a neutral situation when the correlation between the co-occurring IUPAC motifs failed to reach significance. Clustering was carried out according to the degree of their similarity using the Kullback–Leibler distance in the web system STAMP [122].

Figure 5 shows that the motifs have very highly significant ($p < 0.05$) correlations between the co-occurrences of IUPAC motifs in the ChIP-seq peak sequences. Moreover, both significant positive and negative correlations are observed for their co-localization with each other. Clustering according to the degree of similarity in the IUPAC motifs using Kullback–Leibler distance in the web system STAMP [120] showed that all the motifs fall into two large classes. Annotation (see Section 4) of IUPAC motifs using the web resource Tomtom [123] showed that the vast majority of the motifs in Group 1 had a significant similarity with known PWMs for the binding sites of the transcription factor FOXA2. The vast majority of the motifs in Group 2 had a significant similarity with the PWMs of the binding sites for partner transcription factors. As can be seen from the heat map, Group 2 motifs have a subgroup of significantly often co-occurring motifs, but the correlations of their co-occurrence with the motifs of the FOXA2 binding sites are negative. Tomtom

annotation showed that these motifs may significantly ($p < 0.001$) correspond to the binding sites of the transcription factor SP1.

Analysis of pairwise correlations of the co-localization of IUPAC motifs from Group 1 demonstrates a substantial prevalence of positive correlations. At the same time, a small number of negative correlations are also observed.

All this points to complex relationships between TFs in the regulatory regions of genes and a high heterogeneity of data being analyzed. In addition, it was shown that the IUPAC motifs (partner motifs) corresponding to the binding sites of the partner TFs can both positively and negatively correlate with IUPAC motifs (target motifs) corresponding to the binding sites of the target TF FOXA2.

3.2.2. Assessment of the Dependence of the Peak Scores of Peak Sequences on the Presence of Only the Most Significant (P_{Bonf} $(n,N) < 10^{-30}$) Target Motifs in Them

To assess the contribution of the target motifs to the efficiency of peak score prediction, we built a multiple regression model with the use of only the most significant (P_{Bonf} $(n,N) < 10^{-30}$) target motifs corresponding to the binding sites of the target transcription factor FOXA2. To this end, the annotation of the most significant ($p < 10^{-30}$) IUPAC motifs obtained from analysis of ChIP-seq peak sequences in the experiment with FOXA2 was carried out using the web system Tomtom [123] (Section 4), and the motifs that had a significant similarity ($p < 0.001$) with the FOXA2 binding sites PWM were identified. As was found, out of the 50 IUPAC motifs considered, only 22 (44%) were the target motifs. Information about the localization of these motifs in the ChIP-seq peak sequences of the training set **FOXA2$_{\mathbf{L}}$** was used to build a multiple regression model predicting the peak scores of these sequences. Table 8 shows examples of independent variables with significant ($p < 0.05$) regression coefficients and their characteristics. A complete table of the regression coefficients for all IUPAC motifs considered is provided in the Supplementary Materials, Table S4.

Table 8. Examples of independent variables with significant ($p < 0.05$) regression coefficients obtained when constructing a multiple regression model to assess the dependence of the *ln(PS)* value in the ChIP-seq experiment with the transcription factor FOXA2 on the presence of the most significant (P_{Bonf} $(n,N) < 10^{-30}$) target IUPAC motifs. Multiple regression coefficients were calculated using STATISTICA (StatSoft^{TM}, Tulsa, OK, USA).

Independent Variable	Regression Coefficient	*p*-Value
TRTWKACH	0.055717	0.000001
RTTKACHY	0.027913	0.02056
TMAAYANS	0.047318	0.000001
TWKACHYW	0.019684	0.04549
TTKRTYTW	−0.02034	0.046044
KACDTWGN	0.020031	0.035274
TAAHYABW	−0.036249	0.000121
ARMYAAGV	0.019283	0.037902
AAAMAAAR	−0.009348	0.024866

Table 8 shows that nine (41%) of the twenty-two independent variables considered had significant ($p < 0.05$) regression coefficients.

This multiple regression model was used to predict the peak scores of the sequences for the training set **FOXA2$_{\mathbf{L}}$** and the control set **FOXA2$_{\mathbf{C}}$**. Figure 6 shows the observed $ln(PS)_{Obs}$ value for the ChIP-seq peak sequences of the training set **FOXA2$_{\mathbf{L}}$** (Figure 6a) and the control set **FOXA2$_{\mathbf{C}}$** (Figure 6b) versus their expected $ln(PS)_{Exp}$ values calculated with the multiple regression model using information about the localization of the 22 most significant (P_{Bonf} $(n,N) < 10^{-30}$) target motifs.

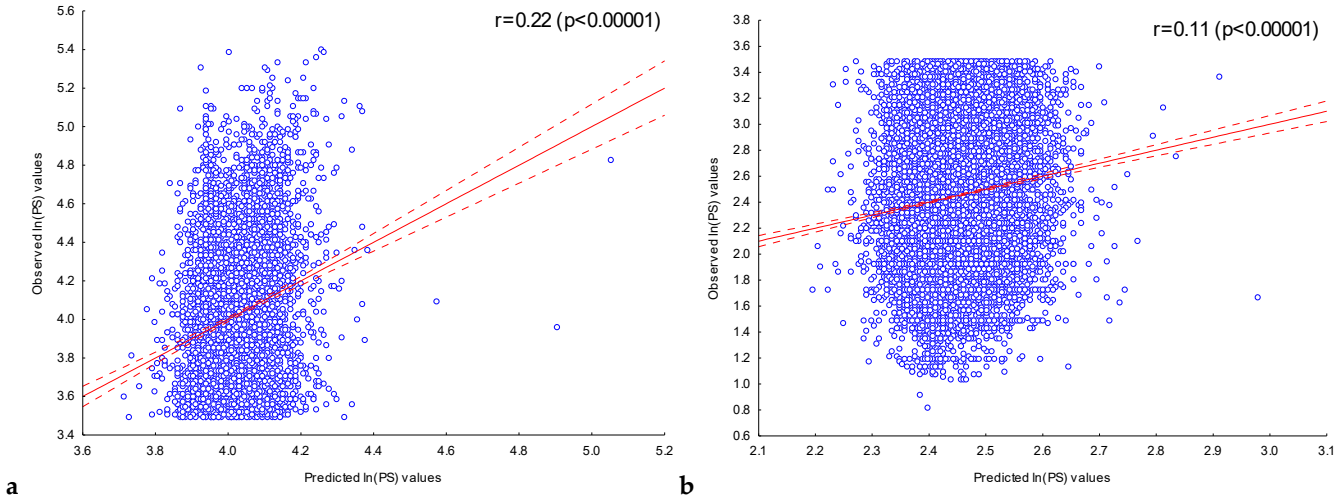


Figure 6. Dependence of the observed values of $ln(PS)_{Obs}$ for ChIP-seq peak sequences from (**a**) the training set **FOXA2$_{L}$** and (**b**) the control set **FOXA2$_{C}$** on their expected values. The expected value of $ln(PS)_{Exp}$ for ChIP-seq peaks was predicted using a multiple regression model for the presence of the most significant (P_{Bonf} $(n,N) < 10^{-30}$) target motifs previously detected in the training set **FOXA2$_{L}$**. Solid and dashed lines represent the regression line and the bounds of its 95% confidence interval as calculated using STATISTICA (StatSoft^{TM}, Tulsa, OK, USA). r is the linear correlation coefficient and p is its statistical significance.

For both sets, the observed $ln(PS)_{Obs}$ and expected $ln(PS)_{Exp}$ values were significantly ($p < 10^{-5}$) correlated. In the case of the training set **FOXA2$_{L}$**, the use of only the 22 most significant (P_{Bonf} $(n,N) < 10^{-30}$) target motifs led to a slightly lower correlation (r= 0.22) than the use of all 50 of the most significant IUPAC motifs, r = 0.24 (Figure 3a). In the case of the control set **FOXA2$_{C}$**, the use of information about the localization of only the 22 most significant (P_{Bonf} $(n,N) < 10^{-30}$) target motifs led to a slightly increased correlation coefficient (r = 0.11) compared to that obtained from using all 50 of the most significant motifs (r = 0.1) (Figure 3b).

Thus, it can be concluded that the use of complete information about the localization of peaks in ChIP-seq sequences of all previously identified significant IUPAC motifs is most effective in predicting the peak scores of ChIP-seq peak sequences in the training sets. Only using information about the presence of the most significant target motifs makes it possible to slightly improve the quality of peak score prediction for ChIP-seq peak sequences in the control sets.

3.2.3. Analysis of the Correlations between the Presence of Target Motifs and the Motifs of the Partner Transcription Factors in the ChIP-Seq Peak Sequences

As can be seen from the analysis of the results obtained, the target motifs make a significant contribution to the prediction of ChIP-seq peak scores; however, sole reliance on target motifs leads to a decrease in prediction quality for ChIP-seq peak sequences that have the highest peak scores and are included in the training sets. To identify transcription factors that can be partners to the target transcription factor FOXA2 in this set and contribute to its functioning when it interacts with the ChIP-seq peak sequences in **FOXA2$_{L}$**, we performed a functional annotation of 28 most significant (P_{Bonf} $(n,N) < 10^{-30}$) IUPAC motifs previously detected in this set and not targeted. The annotation was carried out using the web system Tomtom [123] as described in the Section 4. The results of the annotation are shown in Table 9.

Table 9. Results of the annotation of partner motifs detected in the ChIP-seq peak sequences from the **FOXA2$_L$** set.

Partner TF [1]	Number of Motifs [2]	TF Family [3]	TF Subfamily [4]
HNF4G	6	2.1.3. RXR-related receptors (NR2)	2.1.3.2. HNF-4 (NR2A)
SP1	6	2.3.1. Three-zinc finger Krüppel-related factors	2.3.1.1. SP1-like factors
FOXO3	2	3.3.1. Forkhead box (FOX) factors	3.3.1.15. FOXO
POU5F1	2	3.1.10. POU domain factors	3.1.10.5. POU5 (OCT-3/4-like factors)
NR2F1	1	2.1.3. RXR-related receptors (NR2)	2.1.3.5. COUP-like receptors (NR2F)
ZNF148	1	2.3.3. More than 3 adjacent zinc finger factors	2.3.3.13. ZNF148-like factors
EGR1	1	2.3.1. Three-zinc finger Krüppel-related factors	2.3.1.3. EGR factors
NFYC	1	4.2.1. Heteromeric CCAAT-binding factors	4.2.1.0.3. NF-YC
SOX2	1	4.1.1. SOX-related factors	4.1.1.2. Group B
ZBTB14	1	2.3.3. More than 3 adjacent zinc finger factors	2.3.3.0. unclassified
MAFG	1	1.1.3. MAF-related factors	1.1.3.2. Small Maf factors

[1] Name of the partner transcription factor; [2] Number of motifs significantly ($p < 0.001$) similar to the position–weight matrix of the transcription factor in the JASPAR and HOCOMOCO databases, according to the Tomtom [123] annotation; [3] Number and name of the family to which the TF belongs, according to Wingender's classification [99]; [4] Number and name of the subfamily to which the TF belongs, according to Wingender's classification [99].

We assessed how the motifs corresponding to the binding sites of the partner TF and the motifs corresponding to the binding sites of the target TF FOXA2 co-occur in the **FOXA2$_L$** ChIP-seq peak sequences. Here, in our opinion, the ChIP-seq peak sequence contains the binding site for the TF considered if at least one of the motifs that are significantly similar to the position–weight matrix of this TF is located in this sequence. Consequently, the ChIP-seq peak sequence does not contain a TF binding site if none of the motifs which are significantly similar to the position–weight matrix of this TF are located in the sequence. To exclude from consideration the correlations accounted for by the heterogeneity of the **FOXA2$_L$** set and the peculiarities of its nucleotide context, we assessed the significance of the IUPAC motifs detected and the correlations calculated on their basis using the shuffling procedure (see the Section 4). Table 10 shows the correlation coefficients obtained using the phi-coefficient between the motifs of the partner TF and the motifs of the target TF for the given FOXA2 set.

Table 10. Correlation coefficients obtained using the phi-coefficient to assess the interdependencies of the localization of the motifs corresponding to the binding sites of the partner transcription factors and the motifs corresponding to the binding sites of the target TF FOXA2 on the training set **FOXA2$_L$** of ChIP-seq peak sequences. Significant ($p < 0.05$) correlation coefficients are asterisked.

Partner TF	Correlation Coefficient, r
HNF4G	0.116739 *
SP1	−0.12075
FOXO3	0.106967
POU5F1	0.076699
NR2F1	−0.02758
ZNF148	−0.06076
MAFG	0.106967 *

* Significant ($p < 0.05$) correlation coefficients.

As was found, only two partner transcription factors, HNF4G and MAFG, were characterized by a significant ($p < 10^{-5}$) correlation of the co-occurrence of their potential binding sites with the potential binding sites for FOXA2. Thus, these TFs can potentially interact with FOXA2 in the regions of ChIP-seq peaks and function synergistically with it.

3.2.4. Experimental Data Confirming the Functional Relevance of the Associations Revealed between the Target and Partner Motifs

We wanted to assess the extent to which the information on the co-occurrence of potential binding sites for partner TFs obtained using our proposed approaches is consistent with the experimental data obtained previously and described in the literature. To this end, we analyzed scientific publications describing experimental studies of synergistic interactions with FOXA2. The analysis revealed a number of reports providing experimental evidence for the significant associations that we have found between the target motifs of FOXA2 and the partner motifs of HNF4G in the ChIP-seq peak sequences (Table 11).

Table 11. Experimental evidence for functional interactions between FOXA2 and other partner transcription factors.

Partner TF	Gene or Genomic Regions	Summary	Reference
HNF4 subfamily	Human APOB enhancer	Cooperative interaction between FOXA2 and HNF4 in mediating enhancer function	[124]
HNF4 subfamily (HNF4A)	DNA regions from HepG2 ChIP-seq for FOXA2 and HNF4A	ChIP-sequencing revealed that FOXA2 peaks were co-localizing with HNF4A peaks	[125]
HNF4 subfamily (HNF4A)	DNA regions from adult mouse liver or embryonic hepatoblasts ChIP-seq for FOXA2 and HNF4A	ChIP-sequencing revealed that FOXA2 peaks were co-localizing with HNF4A peaks	[126]
HNF4 subfamily (HNF4A)	Human F2 enhancer	FOXA2 and HNF4A were found to be bound to the enhancer of this gene	[127]
HNF4 subfamily (HNF4A)	DNA regions from adult mouse liver ChIP-seq for FOXA2 and HNF4A	ChIP-sequencing revealed that FOXA2 peaks were co-localizing with HNF4A peaks	[128]

In particular, there is evidence of molecular interactions between FOXA2 and factors from the HNF4 subfamily at human and mouse regulatory DNA elements: (1) co-transfection experiments with the expression vectors for HNF3beta and HNF4 revealed that these factors may bind to the enhancer of the human gene APOB and act synergistically to enhance the intestinal expression of APOB [124]; (2) analyzing DNA regions from ChIP-seq experiments both for FOXA2 and for HNF4a, Wallerman co-workers found that almost half of the FOXA2 peaks were co-localizing with HNF4A peaks, often at a very close distance and with both motifs present [125]; (3) using RNA-seq and ChIP-seq libraries generated from embryonic hepatoblasts and adult mice liver, Alder co-workers showed that the key hepatic TFs HNF4A and FOXA2 occupy enhancers and control target gene expression in a differentiation-dependent manner [126]; (4) in addition, Ceelie co-workers showed that the co-binding of FOXA2 and HNF4A, as well as SP1, with the human prothrombin (F2) enhancer is necessary to ensure an appropriate level of prothrombin expression [127]; (5) analyzing DNA regions from ChIP-seq experiments for FOXA2 and HNF4A in the adult mouse liver, Hoffman co-workers demonstrated FOXA2 and HNF4A collaborations in maintaining the expression of genes that were initially co-expressed in the developing pancreas and liver. They identified 3236 loci in the liver that were co-bound by FOXA2 and HNF4A [128].

Unfortunately, we have not yet found scientific publications that experimentally confirm the synergistic interactions between MAFG and FOXA2. Thus, MAFG may be used as a target for the experimental study of its possible synergistic interactions with FOXA2.

4. Methods and Materials

4.1. Brief Description of the De Novo Motif Discovery System Argo_CUDA

4.1.1. Description of the Criterion Used for the Significance of IUPAC Motifs Detected by Argo_CUDA

The purpose of Argo_CUDA is to iteratively identify significant IUPAC motifs of a fixed length k = 8 bp in a **Pos** set consisting of N sequences of interest of length L. A specific IUPAC motif is considered significant if it is overrepresented in the given set of sequences in **Pos**, its expected presence in **Pos** for random reasons is low, and the probability of its observation in **Pos** for random reasons is significantly low; that is, condition (1) is satisfied.

- $F > f_0$
- $Q < q_0$ (1)
- $P_{Bonf}(n,N) < p_0$

Here $F = n/N$ is the abundance of the motif in **Pos**; that is, the proportion of **Pos** sequences in which the motif occurs at least once; n is the number of **Pos** sequences in which the motif occurs at least once; Q is the expected abundance of the motif in **Pos**; that is, the proportion of **Pos** sequences in which the motif is expected to appear at least once for random reasons; the binomial probability $P_{Bonf}(n,N)$ of observing a motif in at least n of N sequences of the given set **Pos** for random reasons. $P_{Bonf}(n,N)$ is calculated taking into account the Bonferroni correction (see Supplementary Materials). f_0, q_0 and p_0 are user-selectable limit values.

4.1.2. Brief Description of the Argo_CUDA Algorithm

Figure 7 shows a block diagram for our proposed algorithm. This de novo motif discovery algorithm is exhaustive and estimates the significance of the abundance of all 15^k possible IUPAC motifs of fixed length k in the given set **Pos**. Unlike the enumeration and probabilistic approaches, Argo_CUDA can reliably find the global optimum and identify the most significant motif.

In step (1), the given set **Pos** containing the nucleotide sequences of ChIP-seq peaks is fed to Argo_CUDA. In step (2), the frequency characteristics of the set **Pos** are estimated taking into account the Markov level (up to the 2nd order). At the request of the user, the Bernoulli model with a Markov level of 0 and equal frequencies of nucleotides can be used ($P_A = P_T = P_G = P_C = 0.25$). In step (3), Argo_CUDA converts the **Pos** sequences into an array of hashes corresponding to oligonucleotides of length 8 (Table 12). For example, the oligonucleotide **atataaaa** can be represented as a 4-byte binary hash 0001 0010 0001 0010 0001 0001 0001 0001. Thus, each nucleotide sequence of length L is converted into an array of 4-byte hashes of length $L - k + 1$, where k is the length of the oligonucleotide motif.

In step (**4**), for each of the possible 15^k IUPAC motifs, an estimate is made of its expected occurrence Q in **Pos** for random reasons (see Supplementary Materials). It should be noted that a large proportion of all possible motifs may be irrelevant from a biological point of view due to their excessive degeneration (for example, **NNNNNNNN**). There is no need to waste program execution time in assessing the abundance of such motifs in the **Pos** set of DNA sequences, and so such motifs can be eliminated immediately and not considered further. Thus, if for the given motif $Q > q_0$, then according to criterion (1), the motif is excluded from consideration, since it occurs by chance too often. This step significantly narrows the number of possible motifs and substantially speeds up the program. Figure S2 (Supplementary Materials) shows that, for example, at q_0 = 30% in the set of randomly generated sequences, the number of motifs being considered decreases from $15^8 \sim 2.6 \times 10^9$ to ~8.6×10^8, which amounts to about a three-fold decrease. In stage (5), the abundance $F = n/N$ and statistical significance $P_{Bonf}(n,N)$ are calculated for the motifs remaining after filtering. Here, the estimation, n, of the number of **Pos** sequences containing the motif of interest is made using the GPU. In order to substantially accelerate the comparison of the correspondence between DNA regions and IUPAC motifs, we used binary

representations of both motifs and oligonucleotides of DNA sets (Figure 8). A comparison of a motif of length 8 with a DNA region of the same length is performed in one bitwise "AND" operation and one comparison operation instead of eight operations for assessing the correspondence between a nucleotide and a letter of the IUPAC code, which significantly reduces the program runtime: *boolean match = (oligo_hash & motif_hash) = = oligo_hash*.

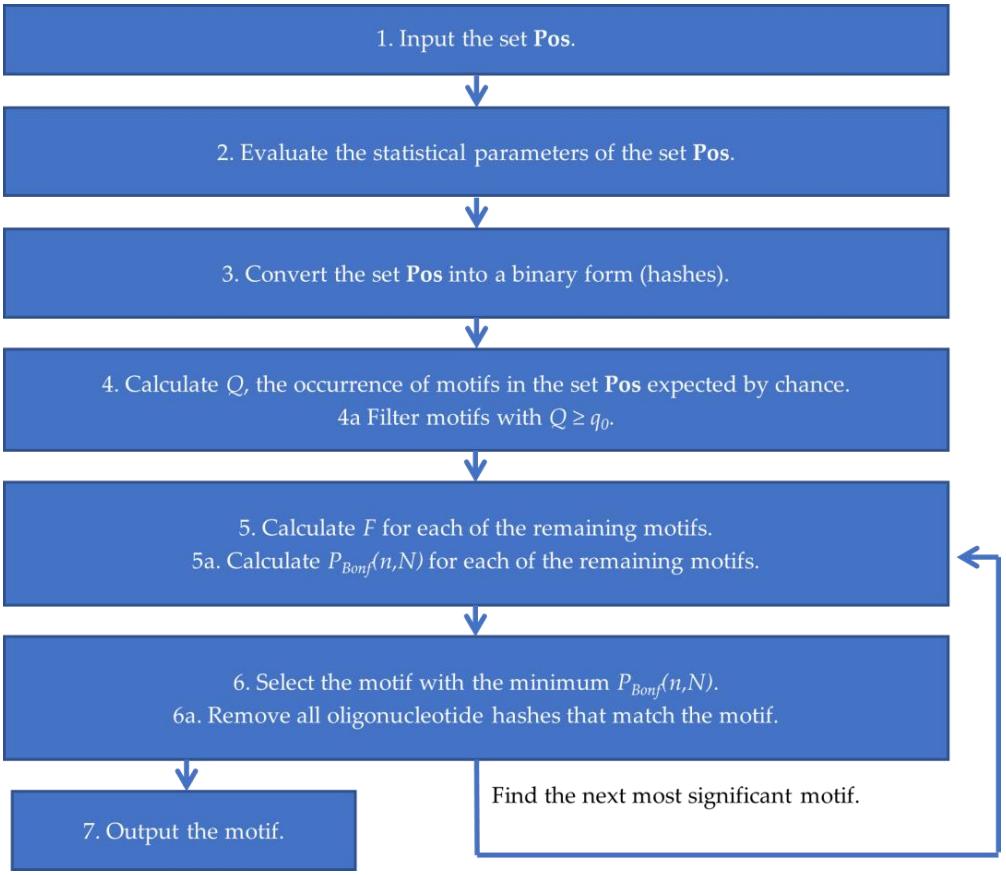


Figure 7. Block diagram for Argo_CUDA, a motif discovery algorithm.

Table 12. Binary representation and hash-coding (hashing) of the 15-letter IUPAC code.

IUPAC Letter	A	T	G	C	R	Y	M	K	W	S	B	H	V	D	N
Nucl. Variants	**A**	**T**	**G**	**C**	**G/A**	**T/C**	**A/C**	**G/T**	**A/T**	**C/G**	**Not A**	**Not G**	**Not T**	**Not C**	**N**
A	1	0	0	0	1	0	1	0	1	0	0	1	1	1	1
T	0	1	0	0	0	1	0	1	1	0	1	1	0	1	1
G	0	0	1	0	1	0	0	1	0	1	1	0	1	1	1
C	0	0	0	1	0	1	1	0	0	1	1	1	1	0	1
Hash	0001	0010	0100	1000	0101	1010	1001	0110	0011	1100	1110	1011	1101	0111	1111

motif_hash	**WWWWAAAA**	0011 0011 0011 0011 0001 0001 0001 0001
oligo_hash	**atataaaa**	0001 0010 0001 0010 0001 0001 0001 0001
boolean match	**true**	0001 0010 0001 0010 0001 0001 0001 0001

Figure 8. An example of comparing the motif **WWWWAAAA** and the oligonucleotide **atataaaa**. Even if any single position of the motif does not match the oligonucleotide, the result of the bitwise "AND" operation is zero.

The proposed optimization methods can substantially speed up the algorithm; however, it will still require a huge amount of computing power. That is why we used high-performance graphics accelerators (GPUs) to explore motif abundance. GPUs are especially powerful for deep parallelization tasks, especially when the data are independent. In step (**6**), the most significant motif with the smallest $P_{Bonf}(n,N)$ that satisfies the significance criterion (1) is selected. This motif is retained, and all the oligonucleotide hashes that correspond to it are removed from **Pos**. Steps 5 and 6 are repeated to identify the next most significant IUPAC motif. After it became impossible to identify any IUPAC motif that satisfies the significance criterion (1) in **Pos**, all previously detected significant motifs at stage (7) are saved to a file and Argo_CUDA stops.

4.1.3. Parameter Values Used to Identify Significant IUPAC Motifs in ChIP-Seq Peak Sequences

Our original development Argo_CUDA was used to identify conserved motifs in the sets of peak sequences obtained from ChIP-seq experiments with ten transcription factors (Table 1). Any motif detected was considered to meet the significance criterion (1) if:

(1) It was located in at least f_0 = 1% of the ChIP-seq peak sequences;
(2) Its expected abundance for random reasons was not more than q_0 = 30%. The expected abundance was calculated taking into account the 3rd order Markov level.
(3) The Bonferroni-corrected binomial probability of observing the motif for random reasons was not more than p_0 = 0.01.

Both DNA strands were examined.

4.2. Building a Multiple Regression Model

The general purpose of a multiple regression approach is to analyze the relationship between several independent variables (also called predictors) and a dependent variable. The objective of a multiple regression analysis is to use independent variables to predict the value of a single dependent variable. The value of each predictor is weighted, and the resulting weights reflect the contribution of each predictor to the overall prediction.

The multiple linear regression equation is

$$Y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_nX_n + e.$$

Here, Y is the dependent variable, $X_1,\ldots,X_n$ are n independent variables, B_0 is the Y-intercept (the value of Y when all other parameters are set to 0), $B_1,\ldots,B_n$ are the regression coefficients of the independent variables $X_1,\ldots,X_n$ and e is the model error. In calculating $B_0,\ldots,B_n$, the regression analysis ensures the maximal prediction of the dependent variable from the set of independent variables and the smallest overall model error e. This calculation is performed using the least squares method.

The construction of a multiple regression model to assess the dependence of peak scores on the presence of significant IUPAC motifs in them for all ten ChIP-seq experiments was carried out using STATISTICA (StatSoftTM, Tulsa, OK, USA) with the "Multiple regression" option.

In this case, the value of the natural logarithm of ChIP-seq peak significance (ln(PS)) was considered as the predicted dependent variable Y, and information about the presence or absence of the corresponding n IUPAC motifs in a particular sequence was considered as the predictors $X_1,\ldots,X_n$. If the i-th motif was present in a particular peak sequence, then X_i = 1, otherwise X_i = 0. For example, for n = 3 motifs, of which only the first and third motifs are present in the j-th sequence of the set-in question, the predicted value is

$$ln(PS)_j = y_j = B_0 + B_1 \times 1 + B_2 \times 0 + B_3 \times 1 = B_0 + B_1 + B_3.$$

For each of the training sets (Table 4) obtained in 10 ChIP-seq experiments, multiple regression models were built on the basis of information about the presence in the peak sequences of this set of n significant IUPAC motifs previously revealed in this set. For each

of the 10 regression models obtained in this way, their predictive power was assessed both on the training sets of peak sequences and on the corresponding control sets (Table 4), by calculating the r values of the correlation between the predicted values of ln(PS) and those observed during the ChIP-seq experiment.

It can be noted that this approach is quite universal, and not only data on the presence/absence of n motifs in sequences but also other information can be used as predictors. Using data from a number of ChIP-seq experiments, we, in particular, assessed the effectiveness of using information about the quantitative representation of each of the motifs, as well as the magnitude of their significance, as predictors. It turned out that these predictors were characterized by slightly worse prediction quality and lower correlation values r on the training and control sets.

4.3. Construction of a Tree of Contextual Similarity of Motifs

A contextual similarity tree of the most significant IUPAC motifs detected in ChIP-seq peak sequences coming from the experiment with the transcription factor FOXA2 was constructed using the web system STAMP [122]. The Kullback–Leibler distance (the "KL" option) was used as a distance measure. The tree was visualized using the MEGA-X system [129].

4.4. Building a Heat Map for the Correlations of the Co-Occurrence of the Motifs

The heat map for the correlations of the co-occurrence of the most significant IUPAC motifs in the ChIP-seq peak sequences obtained from the experiment with the transcription factor FOXA2 was visualized using the web service Heatmapper [121]. Different shades of yellow correspond to different levels of positive correlation; blue, negative correlation; and white, non-significant correlation.

4.5. Functional Annotation of the Motifs

Functional annotation of the most significant IUPAC motifs detected in the ChIP-seq peak sequences coming from the experiment with the target transcription factor FOXA2 was performed using the Tomtom system from the MEME Suite web package [38]. The motifs detected were compared with the PWMs contained in the JASPAR [34] and HOCOMOCO [33] databases with standard parameters. In the first stage of this procedure, the motifs that had significant similarity ($p < 0.001$) with the PWMs of the target transcription factor FOXA2 contained in either of the JASPAR [34] or HOCOMOCO [33] were identified. The IUPAC motifs thus obtained were considered target motifs. The remaining motifs did not have significant similarity with the PWMs of the target TF and were considered partner motifs. For all partner motifs, their similarity with the position–weight matrices of all known TFs contained in the JASPAR [34] and HOCOMOCO [33] databases was assessed. A partner motif was considered to be significantly similar to the transcription factor's position–weight matrix if it had a significant ($p < 0.001$) similarity with the position–weight matrices of this TF in both databases, JASPAR [34] and HOCOMOCO [33].

4.6. Shuffling in Assessing the Significance of the Correlation between the Motifs Corresponding to the Binding Sites of the Target Motifs and the Motifs Corresponding to the Binding Sites of the Partner Transcription Factors

The assessment of the statistical significance of the correlations between the target motifs and partner motifs was carried out using the shuffling procedure. During this procedure, 10,000 random sets were generated by shuffling the nucleotides within each sequence of the given set of ChIP-seq peak sequences. For each generated set of sequences, the abundance, F, of each of the motifs considered was assessed. Next, the motifs, whose abundance F in the set of real ChIP-seq peak sequences exceeded their abundance in 95% of sets of randomly generated sequences, were selected. With the remaining significant IUPAC motifs for each generated random set, the correlation coefficient of the co-occurrence of the motifs corresponding to the binding sites of the target TF and the motifs corresponding to the binding sites of the partner TFs was assessed using the phi-coefficient [120]. Finally,

the proportion of the correlation coefficients that were obtained from random sets and exceeded the absolute value of the correlation coefficient r calculated for the real set of ChIP-seq peak sequences was estimated and taken as the *p*-value.

5. Conclusions

In the present work, we analyzed the contextual organization of ChIP-seq peak sequences in experiments with 10 TFs belonging to the six different superclasses categorized according to the types of their DNA-binding domains [99]. Using the original de novo motif discovery method Argo_CUDA [100], we identified both sets of significant IUPAC motifs corresponding to the target TF binding sites studied in each experiment and specific sets of motifs corresponding to the binding sites of the partner TFs in the peak DNA sequences revealed in each ChIP-seq experiment. Unlike heuristic methods, Argo_CUDA evaluates the significance of all possible IUPAC motifs of a given length, which guarantees finding a global optimum. Our analysis of the ChIP-seq data from the experiment with TF FOXA2 revealed a significant correlation between the presence of the target motifs corresponding to the binding sites for TF FOXA2 and the partner motifs corresponding to the binding sites for TF HNF4G. In the scientific literature, we found experimental evidence for a synergistic interaction between FOXA2 and transcription factors from the HNF4 family, which can explain this correlation. For all the ChIP-seq experiments considered, multiple regression models were constructed, demonstrating a significant dependence of the ChIP-seq peak sequence scores on the presence of sets of specific IUPAC motifs in these sequences. It has been shown that the most significant target motifs make a substantial contribution to the observed dependence. At the same time, the prediction quality can be improved through the use of less significant motifs as well as partner motifs. The contextual features of the ChIP-seq peaks that we have identified can be used to set up experiments aimed at testing potential partner interactions of TFs, the motifs of which are reliably co-represented in the sequences of ChIP-Seq peaks and also help in building potential regulatory gene networks involved in subtle developmental processes and tissue-specific gene expression. In addition, the significant IUPAC motifs we identified can be used to develop new methods for predicting the localization of potential TFBSs in genomic sequences. Unfortunately, despite the fact that we showed a highly reliable dependence of the peak score values on the presence of IUPAC motifs in ChIP-seq sequences, the correlation coefficient r we obtained did not exceed 0.57. This suggests that, to more effectively predict the peak score value, it will be necessary to use additional information, for example, about the relative position of motifs and their orientation in the sequences [130].

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms25021011/s1.

Author Contributions: Conceptualization, O.V.V.; methodology, O.V.V.; software, O.V.V. and A.V.B.; validation, O.V.V.; formal analysis, O.V.V.; investigation, O.V.V. and E.V.I.; resources, O.V.V. and E.V.I.; data curation, O.V.V. and E.V.I.; writing—original draft preparation, O.V.V. and E.V.I.; writing—review and editing, O.V.V. and E.V.I.; visualization, O.V.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Russian government project № FWNR-2022-0020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Links to the analyzed data obtained from the CistromeDB database are given in Table 1. Section 2.1 is devoted to a description of this data.

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

Abbreviations

CE	Composite element
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CRM	Cis-regulatory module
DNA	Deoxyribonucleic acid
GPU	Graphics processing unit
HMM	Hidden Markov model
PS	Peak score
PWM	Position–weight matrix
TF	Transcription factor
TFBS	Transcription factor binding site
TFFM	Transcription factor flexible model

References

1. Weltzien, F.A.; Hildahl, J.; Hodne, K.; Okubo, K.; Haug, T.M. Embryonic development of gonadotrope cells and gonadotropic hormones—Lessons from model fish. *Mol. Cell. Endocrinol.* **2014**, *385*, 18–27. [CrossRef]
2. Wallace, R.M.; Pohler, K.G.; Smith, M.F.; Green, J.A. Placental PAGs: Gene origins, expression patterns, and use as markers of pregnancy. *Reproduction* **2015**, *149*, R115–R126. [CrossRef]
3. Voss, T.C.; Hager, G.L. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet.* **2014**, *15*, 69–81. [CrossRef]
4. Shen, W.K.; Chen, S.Y.; Gan, Z.Q.; Zhang, Y.Z.; Yue, T.; Chen, M.M.; Xue, Y.; Hu, H.; Guo, A.Y. AnimalTFDB 4.0: A comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res.* **2023**, *51*, D39–D45. [CrossRef]
5. Kadonaga, J.T. Eukaryotic transcription: An interlaced network of transcription factors and chromatin-modifying machines. *Cell* **1998**, *92*, 307–313. [CrossRef] [PubMed]
6. Cheng, C.; Alexander, R.; Min, R.; Leng, J.; Yip, K.Y.; Rozowsky, J.; Yan, K.K.; Dong, X.; Djebali, S.; Ruan, Y.; et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* **2012**, *22*, 1658–1667. [CrossRef]
7. Bailey, T.; Krajewski, P.; Ladunga, I.; Lefebvre, C.; Li, Q.; Liu, T.; Madrigal, P.; Taslim, C.; Zhang, J. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.* **2013**, *9*, e1003326. [CrossRef] [PubMed]
8. Collas, P. The current state of chromatin immunoprecipitation. *Mol. Biotechnol.* **2010**, *45*, 87–100. [CrossRef] [PubMed]
9. Johnson, D.S.; Mortazavi, A.; Myers, R.M.; Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **2007**, *316*, 1497–1502. [CrossRef]
10. Park, P.J. ChIP-seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **2009**, *10*, 669–680. [CrossRef]
11. Jakobsen, J.S.; Waage, J.; Rapin, N.; Bisgaard, H.C.; Larsen, F.S.; Porse, B.T. Temporal mapping of CEBPA and CEBPB binding during liver regeneration reveals dynamic occupancy and specific regulatory codes for homeostatic and cell cycle gene batteries. *Genome Res.* **2013**, *23*, 592–603. [CrossRef] [PubMed]
12. Eichenfield, D.Z.; Troutman, T.D.; Link, V.M.; Lam, M.T.; Cho, H.; Gosselin, D.; Spann, N.J.; Lesch, H.P.; Tao, J.; Muto, J.; et al. Tissue damage drives co-localization of NF-κB, Smad3, and Nrf2 to direct Rev-erb sensitive wound repair in mouse macrophages. *eLife* **2016**, *5*, e13024. [CrossRef] [PubMed]
13. Gilmour, J.; Assi, S.A.; Jaegle, U.; Kulu, D.; van de Werken, H.; Clarke, D.; Westhead, D.R.; Philipsen, S.; Bonifer, C. A crucial role for the ubiquitously expressed transcription factor Sp1 at early stages of hematopoietic specification. *Development* **2014**, *141*, 2391–2401. [CrossRef]
14. Oldfield, A.J.; Yang, P.; Conway, A.E.; Cinghu, S.; Freudenberg, J.M.; Yellaboina, S.; Jothi, R. Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors. *Mol. Cell* **2014**, *55*, 708–722. [CrossRef]
15. Sebastian, S.; Faralli, H.; Yao, Z.; Rakopoulos, P.; Palii, C.; Cao, Y.; Singh, K.; Liu, Q.C.; Chu, A.; Aziz, A.; et al. Tissue-specific splicing of a ubiquitously expressed transcription factor is essential for muscle differentiation. *Genes Dev.* **2013**, *27*, 1247–1259. [CrossRef]
16. Wei, H.; Cheng, Y.; Sun, Y.; Zhang, X.; He, H.; Liu, J. Genome-Wide Identification of the *ARF* Gene Family and ARF3 Target Genes Regulating Ovary Initiation in Hazel via ChIP Sequencing. *Front. Plant Sci.* **2021**, *12*, 715820. [CrossRef]
17. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for functional genomics data sets–update. *Nucleic Acids Res.* **2013**, *41*, D991–D995. [CrossRef]
18. Parkinson, H.; Kapushesky, M.; Shojatalab, M.; Abeygunawardena, N.; Coulson, R.; Farne, A.; Holloway, E.; Kolesnykov, N.; Lilja, P.; Lukk, M.; et al. ArrayExpress—A public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* **2007**, *35*, D747–D750. [CrossRef]
19. Burgin, J.; Ahamed, A.; Cummins, C.; Devraj, R.; Gueye, K.; Gupta, D.; Gupta, V.; Haseeb, M.; Ihsan, M.; Ivanov, E.; et al. The European Nucleotide Archive in 2022. *Nucleic Acids Res.* **2023**, *51*, D121–D125. [CrossRef] [PubMed]
20. Kodama, Y.; Shumway, M.; Leinonen, R. International Nucleotide Sequence Database Collaboration The Sequence Read Archive: Explosive growth of sequencing data. *Nucleic Acids Res.* **2012**, *40*, D54–D56. [CrossRef]

21. Nakato, R.; Shirahige, K. Recent advances in ChIP-seq analysis: From quality management to whole-genome annotation. *Brief. Bioinform.* **2017**, *18*, 279–290. [CrossRef] [PubMed]
22. Suryatenggara, J.; Yong, K.J.; Tenen, D.E.; Tenen, D.G.; Bassal, M.A. ChIP-AP: An integrated analysis pipeline for unbiased ChIP-seq analysis. *Brief. Bioinform.* **2022**, *23*, bbab537. [CrossRef] [PubMed]
23. Jeon, H.; Lee, H.; Kang, B.; Jang, I.; Roh, T.Y. Comparative analysis of commonly used peak calling programs for ChIP-Seq analysis. *Genom. Inform.* **2020**, *18*, e42. [CrossRef]
24. Sánchez-Castillo, M.; Ruau, D.; Wilkinson, A.C.; Ng, F.S.; Hannah, R.; Diamanti, E.; Lombard, P.; Wilson, N.K.; Gottgens, B. CODEX: A next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.* **2015**, *43*, D1117–D1123. [CrossRef]
25. Chacon, D.; Beck, D.; Perera, D.; Wong, J.W.; Pimanda, J.E. BloodChIP: A database of comparative genome-wide transcription factor binding profiles in human blood cells. *Nucleic Acids Res.* **2014**, *42*, D172–D177. [CrossRef]
26. Chen, L.; Wu, G.; Ji, H. hmChIP: A database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. *Bioinformatics* **2011**, *27*, 1447–1448. [CrossRef]
27. Zheng, R.; Wan, C.; Mei, S.; Qin, Q.; Wu, Q.; Sun, H.; Chen, C.H.; Brown, M.; Zhang, X.; Meyer, C.A.; et al. Cistrome Data Browser: Expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.* **2019**, *47*, D729–D735. [CrossRef]
28. Kolmykov, S.; Yevshin, I.; Kulyashov, M.; Sharipov, R.; Kondrakhin, Y.; Makeev, V.J.; Kulakovskiy, I.V.; Kel, A.; Kolpakov, F. GTRD: An integrated view of transcription regulation. *Nucleic Acids Res.* **2021**, *49*, D104–D111. [CrossRef]
29. Zou, Z.; Ohta, T.; Miura, F.; Oki, S. ChIP-Atlas 2021 update: A data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res.* **2022**, *50*, W175–W182. [CrossRef] [PubMed]
30. Chen, D.; Jiang, S.; Ma, X.; Li, F. TFBSbank: A platform to dissect the big data of protein-DNA interaction in human and model species. *Nucleic Acids Res.* **2017**, *45*, D151–D157. [CrossRef]
31. Pratt, H.E.; Andrews, G.R.; Phalke, N.; Purcaro, M.J.; van der Velde, A.; Moore, J.E.; Weng, Z. Factorbook: An updated catalog of transcription factor motifs and candidate regulatory motif sites. *Nucleic Acids Res.* **2022**, *50*, D141–D149. [CrossRef]
32. Kaboord, B.; Perr, M. Isolation of proteins and protein complexes by immunoprecipitation. *Methods Mol. Biol.* **2008**, *424*, 349–364. [CrossRef]
33. 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: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **2018**, *46*, D252–D259. [CrossRef] [PubMed]
34. Castro-Mondragon, J.A.; Riudavets-Puig, R.; Rauluseviciute, I.; Lemma, R.B.; Turchi, L.; Blanc-Mathieu, R.; Lucas, J.; Boddie, P.; Khan, A.; Manosalva Pérez, N.; et al. JASPAR 2022: The 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **2022**, *50*, D165–D173. [CrossRef] [PubMed]
35. Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* **2008**, *9*, 326–332. [CrossRef] [PubMed]
36. 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. [CrossRef] [PubMed]
37. Kel, A.E.; Gössling, E.; Reuter, I.; Cheremushkin, E.; Kel-Margoulis, O.V.; Wingender, E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **2003**, *31*, 3576–3579. [CrossRef]
38. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202–W208. [CrossRef]
39. Benos, P.V.; Bulyk, M.L.; Stormo, G.D. Additivity in protein-DNA interactions: How good an approximation is it? *Nucleic Acids Res.* **2002**, *30*, 4442–4451. [CrossRef]
40. Keilwagen, J.; Grau, J. Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.* **2015**, *43*, e119. [CrossRef] [PubMed]
41. Kulakovskiy, I.; Levitsky, V.; Oshchepkov, D.; Bryzgalov, L.; Vorontsov, I.; Makeev, V. From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.* **2013**, *11*, 1340004. [CrossRef] [PubMed]
42. Xu, D.; Liu, H.J.; Wang, Y.F. BSS-HMM3s: An improved HMM method for identifying transcription factor binding sites. *DNA Seq. J. DNA Seq. Mapp.* **2005**, *16*, 403–411. [CrossRef]
43. Wu, J.; Xie, J. Hidden Markov model and its applications in motif findings. *Methods Mol. Biol.* **2010**, *620*, 405–416. [CrossRef]
44. Mathelier, A.; Wasserman, W.W. The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.* **2013**, *9*, e1003214. [CrossRef] [PubMed]
45. Machanick, P.; Bailey, T.L. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* **2011**, *27*, 1696–1697. [CrossRef]
46. Bailey, T.L. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **2011**, *27*, 1653–1659. [CrossRef]
47. Jia, C.; Carson, M.B.; Wang, Y.; Lin, Y.; Lu, H. A new exhaustive method and strategy for finding motifs in ChIP-enriched regions. *PLoS ONE* **2014**, *9*, e86044. [CrossRef]
48. Pavesi, G.; Mereghetti, P.; Mauri, G.; Pesole, G. Weeder Web: Discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* **2004**, *32*, W199–W203. [CrossRef] [PubMed]

49. Sharov, A.A.; Ko, M.S. Exhaustive search for over-represented DNA sequence motifs with CisFinder. *DNA Res.* **2009**, *16*, 261–273. [CrossRef]
50. Zhang, Y.; Wang, P.; Yan, M. An Entropy-Based Position Projection Algorithm for Motif Discovery. *BioMed Res. Int.* **2016**, *2016*, 9127474. [CrossRef]
51. Bailey, T.L.; Williams, N.; Misleh, C.; Li, W.W. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **2006**, *34*, W369–W373. [CrossRef] [PubMed]
52. Ma, X.; Kulkarni, A.; Zhang, Z.; Xuan, Z.; Serfling, R.; Zhang, M.Q. A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res.* **2012**, *40*, e50. [CrossRef]
53. Pavesi, G.; Mauri, G.; Pesole, G. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* **2001**, *17*, S207–S214. [CrossRef] [PubMed]
54. Sinha, S.; Tompa, M. YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* **2003**, *31*, 3586–3588. [CrossRef] [PubMed]
55. Thomas-Chollier, M.; Herrmann, C.; Defrance, M.; Sand, O.; Thieffry, D.; van Helden, J. RSAT peak-motifs: Motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.* **2012**, *40*, e31. [CrossRef]
56. van Helden, J.; André, B.; Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **1998**, *281*, 827–842. [CrossRef] [PubMed]
57. Huo, H.; Zhao, Z.; Stojkovic, V.; Liu, L. Optimizing genetic algorithm for motif discovery. *Math. Comput. Model.* **2010**, *52*, 2011–2020. [CrossRef]
58. Karaboga, D.; Aslan, S. A discrete artificial bee colony algorithm for detecting transcription factor binding sites in DNA sequences. *Genet. Mol. Res. GMR* **2016**, *15*, 1–11. [CrossRef] [PubMed]
59. Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **2015**, *33*, 831–838. [CrossRef]
60. 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. [CrossRef]
61. Worsley Hunt, R.; Mathelier, A.; Del Peso, L.; Wasserman, W.W. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genom.* **2014**, *15*, 472. [CrossRef] [PubMed]
62. 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. [CrossRef]
63. Rasskazov, D.; Chadaeva, I.; Sharypova, E.; Zolotareva, K.; Khandaev, B.; Ponomarenko, P.; Podkolodnyy, N.; Tverdokhleb, N.; Vishnevsky, O.; Bogomolov, A.; et al. Plant_SNP_TATA_Z-Tester: A Web Service That Unequivocally Estimates the Impact of Proximal Promoter Mutations on Plant Gene Expression. *Int. J. Mol. Sci.* **2022**, *23*, 8684. [CrossRef] [PubMed]
64. Abe, N.; Dror, I.; Yang, L.; Slattery, M.; Zhou, T.; Bussemaker, H.J.; Rohs, R.; Mann, R.S. Deconvolving the recognition of DNA shape from sequence. *Cell* **2015**, *161*, 307–318. [CrossRef] [PubMed]
65. Yang, L.; Orenstein, Y.; Jolma, A.; Yin, Y.; Taipale, J.; Shamir, R.; Rohs, R. Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.* **2017**, *13*, 910. [CrossRef] [PubMed]
66. Melikhova, A.V.; Anashkina, A.A.; Il'icheva, I.A. Evolutionary Invariant of the Structure of DNA Double Helix in RNAP II Core Promoters. *Int. J. Mol. Sci.* **2022**, *23*, 10873. [CrossRef]
67. Azeem, M.; Jamil, M.K.; Shang, Y. Notes on the Localization of Generalized Hexagonal Cellular Networks. *Mathematics* **2023**, *11*, 844. [CrossRef]
68. Raza, Z.; Akhter, S.; Shang, Y. Expected value of first Zagreb connection index in random cyclooctatetraene chain, random polyphenyls chain, and random chain network. *Front. Chem.* **2023**, *10*, 1067874. [CrossRef]
69. Kim, J.B.; Spotts, G.D.; Halvorsen, Y.D.; Shih, H.M.; Ellenberger, T.; Towle, H.C.; Spiegelman, B.M. Dual DNA binding specificity of ADD1/SREBP1 controlled by a single amino acid in the basic helix-loop-helix domain. *Mol. Cell Biol.* **1995**, *15*, 2582–2588. [CrossRef] [PubMed]
70. Kouzarides, T. Acetylation: A regulatory modification to rival phosphorylation? *EMBO J.* **2000**, *19*, 1176–1179. [CrossRef]
71. Kemper, J.K.; Xiao, Z.; Ponugoti, B.; Miao, J.; Fang, S.; Kanamaluru, D.; Tsang, S.; Wu, S.Y.; Chiang, C.M.; Veenstra, T.D. FXR acetylation is normally dynamically regulated by p300 and SIRT1 but constitutively elevated in metabolic disease states. *Cell Metab* **2009**, *10*, 392–404. [CrossRef]
72. Vuzman, D.; Hoffman, Y.; Levy, Y. Modulating protein-DNA interactions by post-translational modifications at disordered regions. *Pac. Symp. Biocomput.* **2012**, *2012*, 188–199.
73. Ithuralde, R.E.; Turjanski, A.G. Phosphorylation Regulates the Bound Structure of an Intrinsically Disordered Protein: The p53-TAZ2 Case. *PLoS ONE* **2016**, *11*, e0144284. [CrossRef]
74. Näär, A.M.; Beaurang, P.A.; Robinson, K.M.; Oliner, J.D.; Avizonis, D.; Scheek, S.; Zwicker, J.; Kadonaga, J.T.; Tjian, R. Chromatin, TAFs, and a novel multiprotein coactivator are required for synergistic activation by Sp1 and SREBP-1a in vitro. *Genes Dev.* **1998**, *12*, 3020–3031. [CrossRef] [PubMed]
75. Karczewski, K.J.; Tatonetti, N.P.; Landt, S.G.; Yang, X.; Slifer, T.; Altman, R.B.; Snyder, M. Cooperative transcription factor associations discovered using regulatory variation. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 13353–13358. [CrossRef]

76. Agalioti, T.; Lomvardas, S.; Parekh, B.; Yie, J.; Maniatis, T.; Thanos, D. Ordered Recruitment of Chromatin Modifying and General Transcription Factors to the IFN-b Promoter. *Cell* **2000**, *103*, 667–678. [CrossRef] [PubMed]
77. Koli, S.; Mukherjee, A.; Reddy, K.V.R. Retinoic acid triggers c-kit gene expression in spermatogonial stem cells through an enhanceosome constituted between transcription factor binding sites for retinoic acid response element (RARE), spleen focus forming virus proviral integration oncogene (SPFI1) (PU.1) and E26 transformation-specific (ETS). *Reprod. Fertil. Dev.* **2017**, *29*, 521–543. [CrossRef] [PubMed]
78. Dubois-Chevalier, J.; Mazrooei, P.; Lupien, M.; Staels, B.; Lefebvre, P.; Eeckhoute, J. Organizing combinatorial transcription factor recruitment at cis-regulatory modules. *Transcription* **2018**, *9*, 233–239. [CrossRef]
79. Kel-Margoulis, O.V.; Romashchenko, A.G.; Kolchanov, N.A.; Wingender, E.; Kel, A.E. COMPEL: A database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res.* **2000**, *28*, 311–315. [CrossRef]
80. Matys, V.; Kel-Margoulis, O.V.; Fricke, E.; Liebich, I.; Land, S.; Barre-Dirrie, A.; Reuter, I.; Chekmenev, D.; Krull, M.; Hornischer, K.; et al. TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **2006**, *34*, D108–D110. [CrossRef]
81. Kel-Margoulis, O.V.; Kel, A.E.; Reuter, I.; Deineko, I.V.; Wingender, E. TRANSCompel: A database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.* **2002**, *30*, 332–334. [CrossRef]
82. Merika, M.; Orkin, S.H. Functional synergy and physical interactions of the erythroid transcription factor GATA-1 with the Krüppel family proteins Sp1 and EKLF. *Mol. Cell Biol.* **1995**, *15*, 2437–2447. [CrossRef] [PubMed]
83. Agarwal, P.; Verzi, M.P.; Nguyen, T.; Hu, J.; Ehlers, M.L.; McCulley, D.J.; Xu, S.M.; Dodou, E.; Anderson, J.P.; Wei, M.L.; et al. The MADS box transcription factor MEF2C regulates melanocyte development and is a direct transcriptional target and partner of SOX10. *Development* **2011**, *138*, 2555–2565. [CrossRef]
84. Bieli, D.; Kanca, O.; Requena, D.; Hamaratoglu, F.; Gohl, D.; Schedl, P.; Affolter, M.; Slattery, M.; Müller, M.; Estella, C. Establishment of a Developmental Compartment Requires Interactions between Three Synergistic Cis-regulatory Modules. *PLoS Genet.* **2015**, *11*, e1005376. [CrossRef] [PubMed]
85. Santolini, M.; Sakakibara, I.; Gauthier, M.; Ribas-Aulinas, F.; Takahashi, H.; Sawasaki, T.; Mouly, V.; Concordet, J.P.; Defossez, P.A.; Hakim, V.; et al. MyoD reprogramming requires Six1 and Six4 homeoproteins: Genome-wide cis-regulatory module analysis. *Nucleic Acids Res.* **2016**, *44*, 8621–8640. [CrossRef] [PubMed]
86. Nicolás, M.; Noé, V.; Ciudad, C.J. Transcriptional regulation of the human Sp1 gene promoter by the specificity protein (Sp) family members nuclear factor Y (NF-Y) and E2F. *Biochem. J.* **2003**, *371*, 265–275. [CrossRef]
87. Kerschner, J.L.; Gosalia, N.; Leir, S.H.; Harris, A. Chromatin remodeling mediated by the FOXA1/A2 transcription factors activates CFTR expression in intestinal epithelial cells. *Epigenetics* **2014**, *9*, 557–565. [CrossRef]
88. Swift, M.L.; Beishline, K.; Azizkhan-Clifford, J. Sp1-dependent recruitment of the histone acetylase p300 to DSBs facilitates chromatin remodeling and recruitment of the NHEJ repair factor Ku70. *DNA Repair* **2021**, *105*, 103171. [CrossRef]
89. Wang, L.; Wang, E.; Prado Balcazar, J.; Wu, Z.; Xiang, K.; Wang, Y.; Huang, Q.; Negrete, M.; Chen, K.Y.; Li, W.; et al. Chromatin Remodeling of Colorectal Cancer Liver Metastasis is Mediated by an HGF-PU.1-DPP4 Axis. *Adv. Sci.* **2021**, *8*, e2004673. [CrossRef]
90. Lee, J.S.; Galvin, K.M.; Shi, Y. Evidence for physical interaction between the zinc-finger transcription factors YY1 and Sp1. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 6145–6149. [CrossRef]
91. Millevoi, S.; Thion, L.; Joseph, G.; Vossen, C.; Ghisolfi-Nieto, L.; Erard, M. Atypical binding of the neuronal POU protein N-Oct3 to noncanonical DNA targets. Implications for heterodimerization with HNF-3 beta. *Eur. J. Biochem.* **2001**, *268*, 781–791. [CrossRef] [PubMed]
92. Levitsky, V.; Oshchepkov, D.; Zemlyanskaya, E.; Merkulova, T. Asymmetric Conservation within Pairs of Co-Occurred Motifs Mediates Weak Direct Binding of Transcription Factors in ChIP-Seq Data. *Int. J. Mol. Sci.* **2020**, *21*, 6023. [CrossRef]
93. Whitington, T.; Frith, M.C.; Johnson, J.; Bailey, T.L. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res.* **2011**, *39*, e98. [CrossRef] [PubMed]
94. 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*, 8237–8252. [CrossRef] [PubMed]
95. Deyneko, I.V.; Kel, A.E.; Kel-Margoulis, O.V.; Deineko, E.V.; Wingender, E.; Weiss, S. MatrixCatch—A novel tool for the recognition of composite regulatory elements in promoters. *BMC Bioinform.* **2013**, *14*, 241. [CrossRef] [PubMed]
96. Giannopoulou, E.; Elemento, O. Systematic Discovery of Chromatin-Bound Protein Complexes from ChIP-seq Datasets. *Methods Mol. Biol.* **2017**, *1507*, 43–58. [CrossRef]
97. 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. [CrossRef] [PubMed]
98. Jankowski, A.; Prabhakar, S.; Tiuryn, J. TACO: A general-purpose tool for predicting cell-type-specific transcription factor dimers. *BMC Genom.* **2014**, *15*, 208. [CrossRef]
99. 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. [CrossRef]
100. Vishnevsky, O.V.; Bocharnikov, A.V.; Kolchanov, N.A. Argo_CUDA: Exhaustive GPU based approach for motif discovery in large DNA datasets. *J. Bioinform. Comput. Biol.* **2018**, *16*, 1740012. [CrossRef]

101. Cornish-Bowden, A. Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984. *Nucleic Acids Res.* **1985**, *13*, 3021–3030. [CrossRef] [PubMed]
102. Goldstein, I.; Baek, S.; Presman, D.M.; Paakinaho, V.; Swinstead, E.E.; Hager, G.L. Transcription factor assisted loading and enhancer dynamics dictate the hepatic fasting response. *Genome Res.* **2017**, *27*, 427–439. [CrossRef]
103. Yue, F.; Cheng, Y.; Breschi, A.; Vierstra, J.; Wu, W.; Ryba, T.; Sandstrom, R.; Ma, Z.; Davis, C.; Pope, B.D.; et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **2014**, *515*, 355–364. [CrossRef] [PubMed]
104. MacIsaac, K.D.; Lo, K.A.; Gordon, W.; Motola, S.; Mazor, T.; Fraenkel, E. A quantitative model of transcriptional regulation reveals the influence of binding location on expression. *PLoS Comput. Biol.* **2010**, *6*, e1000773. [CrossRef] [PubMed]
105. Kalvisa, A.; Siersbæk, M.S.; Præstholm, S.M.; Christensen, L.J.L.; Nielsen, R.; Stohr, O.; Vettorazzi, S.; Tuckermann, J.; White, M.; Mandrup, S.; et al. Insulin signaling and reduced glucocorticoid receptor activity attenuate postprandial gene expression in liver. *PLoS Biol.* **2018**, *16*, e2006249. [CrossRef] [PubMed]
106. Zhang, Y.; Laz, E.V.; Waxman, D.J. Dynamic, sex-differential STAT5 and BCL6 binding to sex-biased, growth hormone-regulated genes in adult mouse liver. *Mol. Cell. Biol.* **2012**, *32*, 880–896. [CrossRef] [PubMed]
107. Tsukada, J.; Yoshida, Y.; Kominato, Y.; Auron, P.E. The CCAAT/enhancer (C/EBP) family of basic-leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for gene regulation. *Cytokine* **2011**, *54*, 6–19. [CrossRef] [PubMed]
108. Chen, B.; Lu, Y.; Chen, Y.; Cheng, J. The role of Nrf2 in oxidative stress-induced endothelial injuries. *J. Endocrinol* **2015**, *225*, R83–R99. [CrossRef]
109. Kopacz, A.; Kloska, D.; Klimczyk, D.; Kopec, M.; Jozkowicz, A.; Piechota-Polanczyk, A. Nrf2 Transcriptional Activity Governs Intestine Development. *Int. J. Mol. Sci.* **2022**, *23*, 6175. [CrossRef]
110. Barbarani, G.; Fugazza, C.; Strouboulis, J.; Ronchi, A.E. The Pleiotropic Effects of GATA1 and KLF1 in Physiological Erythropoiesis and in Dyserythropoietic Disorders. *Front. Physiol.* **2019**, *10*, 91. [CrossRef]
111. Tachmatzidi, E.C.; Galanopoulou, O.; Talianidis, I. Transcription Control of Liver Development. *Cells* **2021**, *10*, 2026. [CrossRef]
112. Ferdous, A.; Hill, J.A. FoxO1 in embryonic development. *Transcription* **2012**, *3*, 221–225. [CrossRef] [PubMed]
113. Remadevi, V.; Muraleedharan, P.; Sreeja, S. FOXO1: A pivotal pioneer factor in oral squamous cell carcinoma. *Am. J. Cancer Res.* **2021**, *11*, 4700–4710. [PubMed]
114. Basile, V.; Baruffaldi, F.; Dolfini, D.; Belluti, S.; Benatti, P.; Ricci, L.; Artusi, V.; Tagliafico, E.; Mantovani, R.; Molinari, S.; et al. NF-YA splice variants have different roles on muscle differentiation. *Biochim. Biophys. Acta* **2016**, *1859*, 627–638. [CrossRef] [PubMed]
115. Darvin, P.; Joung, Y.H.; Yang, Y.M. JAK2-STAT5B pathway and osteoblast differentiation. *JAKSTAT* **2013**, *2*, e24931. [CrossRef]
116. Gao, P.; Zhang, Y.; Liu, Y.; Chen, J.; Zong, C.; Yu, C.; Cui, S.; Gao, W.; Qin, D.; Sun, W.; et al. Signal transducer and activator of transcription 5B (STAT5B) modulates adipocyte differentiation via MOF. *Cell Signal* **2015**, *27*, 2434–2443. [CrossRef] [PubMed]
117. Georganta, E.M.; Tsoutsi, L.; Gaitanou, M.; Georgoussi, Z. δ-opioid receptor activation leads to neurite outgrowth and neuronal differentiation via a STAT5B-Gαi/o pathway. *J. Neurochem.* **2013**, *127*, 329–341. [CrossRef]
118. Baker, W.; van den Broek, A.; Camon, E.; Hingamp, P.; Sterk, P.; Stoesser, G.; Tuli, M.A. The EMBL nucleotide sequence database. *Nucleic Acids Res.* **2000**, *28*, 19–23. [CrossRef]
119. Bonferroni, C.E. Teoria statistica delle classi e calcolo delle probabilità. *Pubbl. R Ist. Super. Sci. Econ. Commer. Firenze* **1936**, *8*, 1–62.
120. Cramer, H. *Mathematical Methods of Statistics*; Princeton University Press: Princeton, NJ, USA, 1946; p. 282.
121. Babicki, S.; Arndt, D.; Marcu, A.; Liang, Y.; Grant, J.R.; Maciejewski, A.; Wishart, D.S. Heatmapper: Web-enabled heat mapping for all. *Nucleic Acids Res.* **2016**, *44*, W147–W153. [CrossRef]
122. Mahony, S.; Benos, P.V. STAMP: A web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* **2007**, *35*, W253–W258. [CrossRef] [PubMed]
123. Gupta, S.; Stamatoyannopoulos, J.A.; Bailey, T.L.; Noble, W.S. Quantifying similarity between motifs. *Genome Biol.* **2007**, *8*, R24. [CrossRef] [PubMed]
124. Antes, T.J.; Levy-Wilson, B. HNF-3 beta, C/EBP beta, and HNF-4 act in synergy to enhance transcription of the human apolipoprotein B gene in intestinal cells. *DNA Cell Biol.* **2001**, *20*, 67–74. [CrossRef] [PubMed]
125. 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. [CrossRef] [PubMed]
126. Alder, O.; Cullum, R.; Lee, S.; Kan, A.C.; Wei, W.; Yi, Y.; Garside, V.C.; Bilenky, M.; Griffith, M.; Morrissy, A.S.; et al. Hippo signaling influences HNF4A and FOXA2 enhancer switching during hepatocyte differentiation. *Cell Rep.* **2014**, *9*, 261–271. [CrossRef] [PubMed]
127. Ceelie, H.; Spaargaren-Van Riel, C.C.; De Jong, M.; Bertina, R.M.; Vos, H.L. Functional characterization of transcription factor binding sites for HNF1-alpha, HNF3-beta (FOXA2), HNF4-alpha, Sp1 and Sp3 in the human prothrombin gene enhancer. *J. Thromb. Haemost.* **2003**, *1*, 1688–1698. [CrossRef]
128. Hoffman, B.G.; Robertson, G.; Zavaglia, B.; Beach, M.; Cullum, R.; Lee, S.; Soukhatcheva, G.; Li, L.; Wederell, E.D.; Thiessen, N.; et al. Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver. *Genome Res.* **2010**, *20*, 1037–1051. [CrossRef]

129. 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. [CrossRef]
130. Vishnevsky, O.V.; Kolchanov, N.A. ARGO: A web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters. *Nucleic Acids Res.* **2005**, *33* (Suppl. S2), W417–W422. [CrossRef]