Вавиловский журнал генетики и селекции. 2021;25(1):7-17
DOI 10.18699/VJ21.002

Original Russian text www.bionet.nsc.ru/vogis/

Application of alternative *de novo* motif recognition models for analysis of structural heterogeneity of transcription factor binding sites: a case study of FOXA2 binding sites

A.V. Tsukanov[1] ✉, V.G. Levitsky[1, 2], T.I. Merkulova[1, 2]

[1] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
✉ tsukanov@bionet.nsc.ru

Abstract. The most popular model for the search of ChIP-seq data for transcription factor binding sites (TFBS) is the positional weight matrix (PWM). However, this model does not take into account dependencies between nucleotide occurrences in different site positions. Currently, two recently proposed models, BaMM and InMoDe, can do as much. However, application of these models was usually limited only to comparing their recognition accuracies with that of PWMs, while none of the analyses of the co-prediction and relative positioning of hits of different models in peaks has yet been performed. To close this gap, we propose the pipeline called MultiDeNA. This pipeline includes stages of model training, assessing their recognition accuracy, scanning ChIP-seq peaks and their classification based on scan results. We applied our pipeline to 22 ChIP-seq datasets of TF FOXA2 and considered PWM, dinucleotide PWM (diPWM), BaMM and InMoDe models. The combination of these four models allowed a significant increase in the fraction of recognized peaks compared to that for the sole PWM model: the increase was 26.3 %. The BaMM model provided the main contribution to the recognition of sites. Although the major fraction of predicted peaks contained TFBS of different models with coincided positions, the medians of the fraction of peaks containing the predictions of sole models were 1.08, 0.49, 4.15 and 1.73 % for PWM, diPWM, BaMM and InMoDe, respectively. Thus, FOXA2 BSs were not fully described by only a sole model, which indicates theirs heterogeneity. We assume that the BaMM model is the most successful in describing the structure of the FOXA2 BS in ChIP-seq datasets under study.
Key words: transcription factor binding sites (TFBS); TFBS *de novo* searching; ChIP-seq; heterogeneity of TFBS.

For citation: Tsukanov A.V., Levitsky V.G., Merkulova T.I. Application of alternative *de novo* motif recognition models for analysis of structural heterogeneity of transcription factor binding sites: a case study of FOXA2 binding sites. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):7-17. DOI 10.18699/VJ21.002

Метод поиска структурной гетерогенности сайтов связывания транскрипционных факторов с использованием альтернативных *de novo* моделей на примере FOXA2

А.В. Цуканов[1] ✉, В.Г. Левицкий[1, 2], Т.И. Меркулова[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ tsukanov@bionet.nsc.ru

Аннотация. В настоящее время самой распространенной моделью поиска сайтов связывания транскрипционных факторов (ССТФ) в пиках ChIP-seq является позиционная весовая матрица (position weight matrix, PWM). Но эта модель не учитывает взаимосвязи между частотами встреч нуклеотидов в разных позициях ССТФ, поэтому не способна гарантировать определение всех возможных структурных вариантов ССТФ. На сегодняшний день уже предложены альтернативные модели, например BaMM и InMoDe, которые учитывают такие взаимосвязи. Однако применение этих моделей обычно сводилось к сравнению их точности с точностью традиционной модели PWM, тогда как анализ совместной встречаемости и относительного расположения ССТФ разных моделей в пиках не производился. В нашей работе мы предлагаем конвейер программ MultiDeNA, позволяющий сочетать разные модели *de novo* поиска ССТФ для выявления структурной гетерогенности ССТФ в данных ChIP-seq. Разработанный конвейер включает этапы построения моделей на основе заданного набора пиков, оценки точности распознавания моделей с помощью перекрестных тестов, выбора порогов, сканирования пиков ChIP-seq и классификацию пиков по результатам сканирования. С применением конвейера нами проведен анализ 22 экспериментов ChIP-seq для ТФ FOXA2 с помощью четырех моделей: PWM, diPWM, BaMM и InMoDe. Показано, что сочетание моделей позволяет существенно увеличить

общее количество распознанных пиков (на 26.3 %) по сравнению с применением только PWM; при этом основной вклад в распознавание внесла модель BaMM. В значительной доле пиков разные модели распознают совпадающие ССТФ; однако для моделей PWM, diPWM, BaMM и InMoDe медианы доли пиков, которые содержали ССТФ только одной модели, составили 1.08, 0.49, 4.15 и 1.73 % соответственно. Таким образом, совокупность ССТФ FOXA2 не описывается полностью только одной моделью, что свидетельствует о наличии структурной гетерогенности в ССТФ у FOXA2.
Ключевые слова: сайты связывания транскрипционных факторов (ССТФ); *de novo* поиск ССТФ; ChIP-seq; гетерогенность ССТФ.

Introduction

Transcription factors (TFs) are proteins that can recognize certain regions of genomic DNA (TF binding sites, TFBS) (Lambert et al., 2018). The main function of TFs is to increase or decrease a level of gene transcription (Latchman, 2001). The key stage of the regulation of gene expression is TF binding to DNA. This binding initiates a chain of molecular events that ensure the assembly and regulate the activity of the pre-initiation complex of RNA polymerase II, both through direct or indirect contacts with the components of this complex, and through the involvement of various modifying chromatin and remodeling proteins. As a consequence, local changes in the structure of chromatin allow the transcription initiation (Iwafuchi-Doi, 2019; Srivastava, Mahony, 2020). Therefore, one of the most important tasks of modern molecular biology is to identify genomic TFBSs.

Currently, the ChIP-seq technique is widely used to solve this problem (Farnham, 2009; Park, 2009). This technique is based on the chromatin immunoprecipitation with antibodies to an investigated TF with consequent high-throughput sequencing of precipitated DNA. Primary ChIP-seq data processing identifies DNA regions, or peaks, in which a target TF was directly or through intermediate proteins binds DNA (Furey, 2012). However, lengths of peaks are usually equal to hundreds of bp, but a length of TFBS does not exceed 20–25 bp (Levitsky et al., 2007; Kulakovskiy et al., 2018). Thus, the next stage of the bioinformatics processing of ChIP-seq data is to search exact positions of TFBS in peaks. To date, many tools have been developed to solve this issue, the overwhelming majority of them are based on the model of position weight matrix (PWM) (Stormo, 2000), including such popular ones as ChIPMunk (Kulakovskiy, Makeev, 2009) and Homer (Heinz et al., 2010). It is no exaggeration to say that the use of different implementations of the PWM model are included in almost every pipeline of ChIP-seq data processing (Lloyd, Bao, 2019).

The application of the standard PWM-based approach to the processing of ChIP-seq data showed that for most TFs about a half of peaks did not contain detected PWM hits (Worsley Hunt, Wasserman, 2014; Gheorghe et al., 2019). Traditionally, this was associated with the main disadvantage of PWM, the hypothesis of independence of nucleotides frequencies in different positions of TFBS, which is not always true. This may negatively affect the recognition accuracy (Benos et al., 2002; Keilwagen, Grau, 2015). Therefore, alternative models of TFBS recognition have being developed. They took into account dependencies between nucleotides occurrences in a site model (Mathelier, Wasserman, 2013; Yang et al., 2014; Siebert, Söding, 2016; Eggeling et al., 2017; Gheorghe et al., 2019). Thus, the simplest alternative model was the dinucleotide position weight matrix (diPWM), it took into account dependences between adjacent nucleotides (Zhang M., Marr, 1993; Kulakovskiy et al., 2013). On the other hand, models such as BaMM (Siebert, Söding, 2016) and InMoDe (Eggeling et al., 2017) have been proposed. They were constructed using Markov chains, which took into account dependences of positions using the concept of Markov chain order, i. e. a length for which nucleotide frequencies can be mutually dependent (an order usually does not exceed 5 nt).

Authors of these alternative models proved that their models might outperform in recognition accuracy the standard PWM. However, these models were not applied to solve the problem of incomplete recognition of TFTS in ChIP-seq peaks. We assume that this problem is partially related to the structural heterogeneity of binding sites of TFs, and the number of recognized peaks can be significantly increased with the combination of different models together. In this case, the ChIP-seq peaks contain both predicted TFBS with application of a sole model, or with two models, etc. (Ignatieva et al., 2004; Levitsky et al., 2014, 2016). Earlier, we used the training sample of 53 known TF sites of the FOXA subfamily and analyzed ChIP-seq data of FOXA2 (Wederell et al., 2008; Wallerman et al., 2009) with alternative models ChIPMunk (PWM) (Kulakovskiy, Makeev, 2009) and SiteGA (Levitsky et al., 2007) with experimentally fitted model's thresholds (EMSA experiment, electrophoretic mobility shift assay, shift in electrophoretic mobility analysis). We showed that both models together found FOXA2 sites in more than 95 % of peaks (Levitsky et al., 2014). This conclusion was consistent with the absence in literature any data about indirect interaction of this well-studied TF with genomic DNA.

The given example indicates that combination of alternative models with PWM model for analyzing ChIP-seq data is promising. However, until now there has been no systematic research on this topic. Alternative models of TFBS search are not widely used, despite that about 20 years ago it was proved that there is a dependence of the nucleotide frequencies in different positions in TFBS (Bulyk et al., 2002). As the indicator of the popularity of different models, we use the number of citations of papers devoted to specific *de novo* TFBS searching programs for ChIP-seq data analysis. Thus, at the end of 2020, papers

devoted to the implementation of the PWM model MEME, HOMER and ChIPMunk (Bailey, Elkan, 1994; Heinz et al., 2010; Kulakovskiy et al., 2010; Machanick, Bailey, 2011) have the total number of citations over 6000. However, papers devoted to alternative models BaMM, InMoDe and diChIPMunk (Kulakovskiy et al., 2013; Siebert, Söding, 2016; Eggeling et al., 2017; Kiesel et al., 2018) have about 50 citations. Moreover, specific studies of individual ChIP-seq experiments were usually analyzed only with the standard PWM model. This situation we explain as follow. First, the PWM model is understandable and anyone can simply interpret it. Second, advantages of alternative models are insufficiently understandable. E. g., hardly anyone thought that alternative models were able systematically to find out TFBS of a different structure.

In this paper, we propose the pipeline that combines four *de novo* models of TFBS search, namely ChIPMunk/diChIPMunk implementations of PWM/diPWM (Kulakovskiy et al., 2010, 2013), and the Markov models InMoDe (Eggeling et al., 2017) and BaMM (Siebert, Söding, 2016). The pipeline evaluates the recognition accuracy of these models, selects their thresholds and classifies ChIP-seq peaks by comparing respective scan results. This approach expands the understanding of TFBS structural diversity, especially in cases when the PWM model is unable to find TFBS in a peak. We applied the pipeline for 22 ChIP-seq datasets for TF FOXA2.

Materials and methods

For the analysis we used the set of preprocessed 22 ChIP-seq datasets for TF FOXA2 in the bed format from the ReMap database http://remap.univ-amu.fr/ (Chèneby et al., 2020), see the Table. Only the best 4000 peaks we used for analysis in each sample (see below).

The input of our pipeline includes a dataset of ChIP-seq peaks with notation of genome version (mm10 or hg38) and the list of available TFBS search programs (PWM, diPWM, BaMM, InMoDe). The notation of genome version allows selection of the list of promoters in the fasta format (5′-regions of protein-coding genes, 2000 bp upstream transcription start sites). This promoter dataset is required for concordant threshold selection for all models. The total sizes of these samples were 19795/19991 genes for the human/mouse genomes (GRCh38.p13/GRCm38.p6 versions). We used the reference genome to extract nucleotide sequences of the peaks.

Pipeline for searching heterogeneity of TFBS. We have developed the MultiDeNA pipeline (multiple de novo analysis, https://github.com/ubercomrade/MultiDeNA) to search TFBS in ChIP-seq data with several *de novo* models. This pipeline allows obtaining the classification of ChIP-seq peaks, which is used to estimate the structural diversity of TFBS. The pipeline currently uses ChIPMunk (PWM), diChIPMunk (diPWM), BaMM, and InMoDe models, as well as the bedtools (Quinlan, Hall, 2010) and TomTom (Gupta et al., 2007) support programs. The schematic diagram of the program pipeline is shown in Fig. 1. The

The list of ChIP-seq experiments used in our study

No.	GEO/ ENCODE ID	Cell line/tissue	Treatment	TomTom
1	ENCSR066EBK	Hep-G2	–	+
2	GSE90454	BJ1-hTERT	Mimosine	+
3	GSE90454	A-549	–	+
4	ENCSR000BRE	A-549	–	+
5	GSE92491	BJ1-hTERT	Mimosine	+
6	GSE90454	BJ1-hTERT	–	+
7	ENCSR080XEY	Liver	–	+
8	ENCSR310NYI	Liver	–	+
9	ENCSR000BNI	Hep-G2	–	+
10	GSE90454	BJ1-hTERT	–	+
11	ERP004206	H9	–	+
12	GSE92491	BJ1-hTERT	Mimosine	–
13	GSE90454	KerCT	–	+
14	GSE90454	BJ1-hTERT	Mimosine	–
15	GSE90454	BJ1-hTERT	Mimosine	+
16	GSE90454	BJ1-hTERT	Mimosine	+
17	GSE90454	BJ1-hTERT	GATA4	–
18	ERP008682	Pancreas	CARN1618	+
19	GSE90454	BJ1-hTERT	Mimosine	–
20	GSE92491	BJ1-hTERT	CDT1	+
21	GSE90454	Hep-G2	–	–
22	GSE92491	BJ1-hTERT	FOXA2 and GATA4 coexpression	–

Note: GEO/ENCODE – unique identifier of databases (GSE*/ENC*). TomTom – result of filtering data using TomTom software (see "Comparison of found TFBS with known ones using TomTom tool"). (+)/(–) – the frequency matrix built on the basis of the TFBS found by ChIPMunk (PWM) is significantly similar (p-value < 0.001)/not similar (p-value > 0.001) to the frequency matrix of the FOXA2 TFBS from HOCOMOCO FOXA2_HUMAN.H11MO.0.A.

pipeline includes the following steps: (1) data preparation, (2) building of a model, (3) model accuracy assessment, (4) threshold selection and search of TFBS in ChIP-seq peaks with the fixed thresholds and (5) classification of ChIP-seq peaks according to results of TFBS recognition. Each stage of the program pipeline is described below.

Preparing initial data for analysis. The preparation of the data included the sorting of peaks according the value $-10 \cdot \log_{10}$ (p-value) that characterized the peak quality. This value was previously calculated for each peak by the MACS program (Zhang Y. et al., 2008). The pipeline of ReMap database (Chèneby et al., 2020) used this program to process raw ChIP-seq data. For each ChIP-seq dataset,

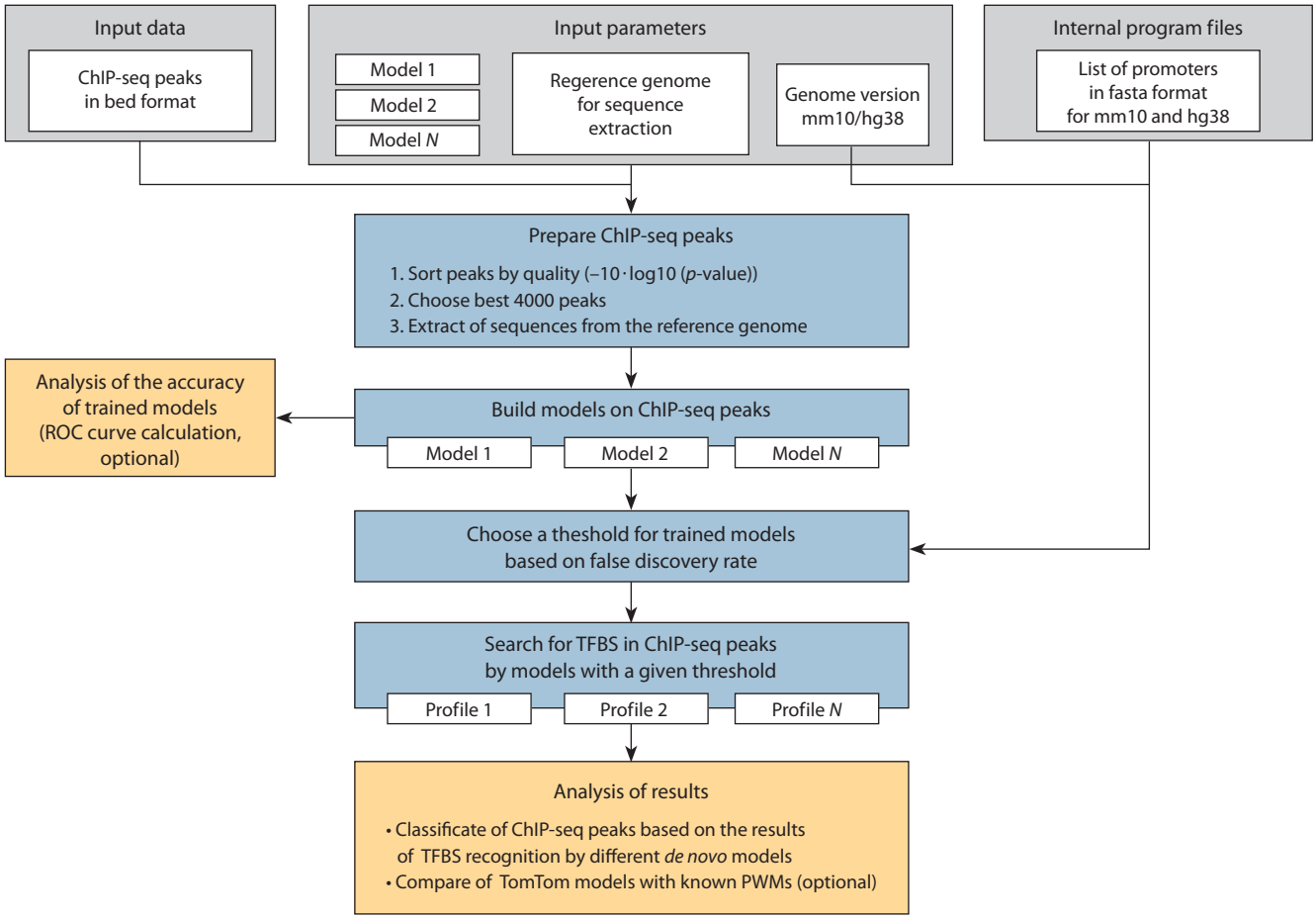


Fig. 1. The scheme of MultiDeNA workflow.

we took in analysis top-scoring 4000 peaks. Next, nucleotide sequences of the peaks we extracted from the genome using bedtools (Quinlan, Hall, 2010).

Training *de novo* models and assessing the TFBS recognition accuracy. In order to recognize TFBS in peaks, it is necessary to build *de novo* models. The PWM and diPWM models we build with ChIPMunk and diChIPMunk, respectively (Kulakovskiy et al., 2010, 2013).The construction of alternative models we carry out with BaMM (Siebert, Söding, 2016) and InMoDe (Eggeling et al., 2017).

To improve the recognition accuracy of PWM model, we selected it optimal length by the cross-validation procedure. We used the same length for the construction of other models. This procedure included the following steps: (1) to divide the ChIP-seq dataset randomly into the training (90 % of the peaks) and the control (remaining 10 % of the peaks) samples; (2) to build a model with the training sample; (3) to get recognition scores of a model with the control sample to calculate true positive rate (TPR); (4) to generate the sample of random sequences by shuffling of nucleotides in the control sample; (5) to get scores of a model with the sample of random sequences to calculate the false positives rate (FPR); (6) repetition of steps 1–5 several times; (7) to calculate the ROC-curve (receiver operating characteristic). We compared different models with the pAUC value (partial area under curve), we calculated it as the part of the area under ROC curve for all FPR values less than 0.001 (McClish, 1989; Siebert, Söding, 2016). The method described above for choosing the optimal PWM length was developed earlier (Levitsky et al., 2007; Kulakovskiy et al., 2013). The accuracy of all models we assessed with the same approach.

Next, a model can be applied to a nucleotide sequence with the same length as a model site. The result of applying this model is the recognition score. The larger score points to the higher probability of estimated nucleotide sequence to be a functional TFBS.

Threshold selection for models according to false positive rate estimates. To compare the results of TFBS search of different models correctly, it is necessary to set thresholds for all models uniformly. We set these thresholds for all models according the fixed FPR. To calculate this FPR, we use the negative sample, which included 5′-regions of protein-coding genes (2000 base pairs from transcription start sites).

We calculate FPR as follows. The scores of a model we determine for each site in the negative sample at each position and DNA strand. Then, the FPR for each unique

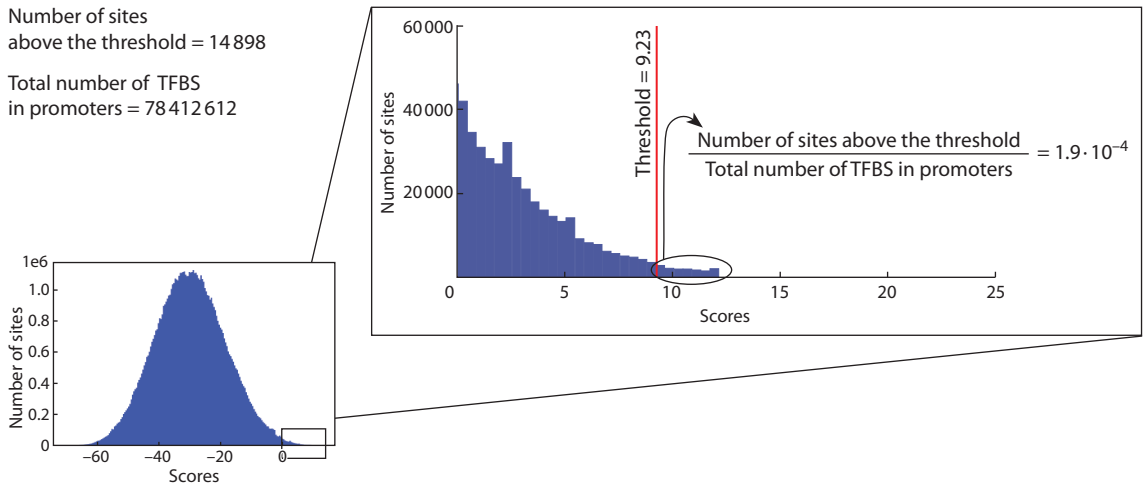


Fig. 2. The approach of threshold selection for a model through estimation of false positive rate with the whole-genome promoter dataset.

score threshold we calculate as the ratio of the total count of predicted TFBS, for which the score is the same or higher than this threshold, to the total number of positions in the sequence sample available for such TFBS. We choose for recognition of TFBS in peaks thresholds for all models respecting the FPR $1.9 \cdot 10^{-4}$. An example of choosing a threshold for PWM for the GSE92491 dataset is shown in Fig. 2.

Classification of ChIP-seq peaks based on the results of TFBS recognition by different models. After threshold selection for all models, we search TFBS in ChIP-seq peaks. Further, these peaks we classify into fractions depending on the presence/absence of sites of different models (PWM, diPWM, BaMM, InMoDe). We use two types of classification. One of them take into account the location of TFBS of different models in a peak, and another did not (see previously developed method, Levitsky et al., 2014, 2016). In particular, we carry out for each pair of models the classification of peaks with taking into account positions of TFBS of different models. Totally, there are six pairs of models: PWM and diPWM, PWM and BaMM, PWM and InMoDe, BaMM and diPWM, BaMM and InMoDe, InMoDe and diPWM. If a peak includes TFBS of a single model only, then this peak we classify as the peak of the corresponding model. If there are only two different models with hits in a peak, then two outcomes are possible (Fig. 3).

In the first case, if there is at least one pair of sites from two models that has at least one common position, then such peak we classify as the "intersection". Otherwise, if a peak contains sites of different models, but these sites are not intersected, then a peak is classified as "no intersection". If sites are absent in a peak, then we classify it as "no sites". Such classification of ChIP-seq peaks for the two models can be represented as the pie chart (Fig. 4).

The classification of peaks, without taking into account positions of sites of different models we carry out as follows. We identify following groups of peaks: peaks with

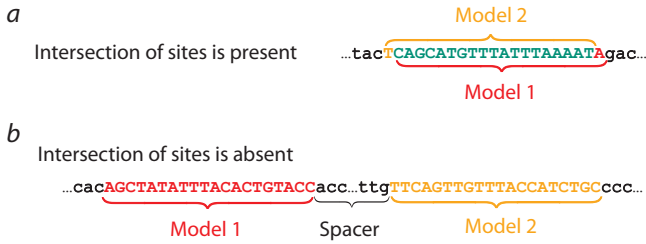


Fig. 3. The example of classification for two ChIP-seq peaks containing sites of two various models (Model 1, Model 2). Colors mark options of intersected (*a*) or not intersected sites (*b*).

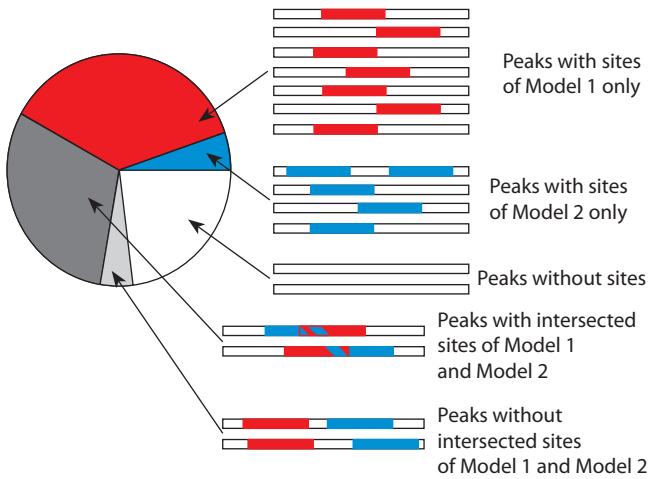


Fig. 4. Peak classification for two models (Model 1, Model 2) with taking into account the intersection of TFBS.

sites of one model only, peaks containing sites of all models, and also several groups of peaks respecting combination of various models.

Comparison of found TFBS with known ones using TomTom tool. To assess whether a predicted site matches to known FOXA2 sites, we use the TomTom motif

comparison program (Gupta et al., 2007). This program is designed to assess the similarity between nucleotide frequency matrices. For each PWM model, we construct a nucleotide frequency matrix based on the sites it find. Next, using TomTom, we evaluate the similarity of this matrix to the frequency matrix of the FOXA2 from the HOCOMOCO database (ID HOCOMOCO FOXA2_HUMAN.H11MO.0.A, Kulakovskiy et al., 2018). If the *p*-value of the matrix comparison is below 0.001, then a ChIP-seq dataset we consider as enriched with FOXA2 BS (see the Table).

Statistical data analysis. Data analysis and visualization we perform in the Python 3.8 programming language in the Jupyter environment using the numpy, matplotlib, seaborn, and statannot packages. The distributions of peak fractions respecting to various models we compare with the Mann–Whitney U-test, corrected for multiple comparisons (Bonferroni approach).

Results

Filtering data based on TomTom's motif comparison

To ensure that the trained models find sites corresponding to known FOXA2 sites we apply the filter based on the TomTom program. For this, we build the frequency matrices respecting a trained model and we compare it with the known matrix of FOXA2 from the HOCOMOCO database. This procedure left only 16 ChIP-seq datasets out of total 22 (see the Table), therefore, these 16 sets we use in further analysis.

Classification of ChIP-seq peaks without taking into account the intersection of TFBS positions found by different *de novo* models

The main result of MultiDeNA pipeline is the classification of peaks. It allows establishing how the models are related to each other in terms of their ability to identify TFBS in peaks. We used two types of peak classification. The first one takes into account an intersection of positions of predicted TFBS of different models, the second one did not take it into account (see "Classification of ChIP-seq peaks based on the results of TFBS recognition by different models"). The example of results classification for GSE90454.FOXA2.KerCT dataset is given in Fig. 5.

Let us consider in more detail the classification of ChIP-seq peaks based on the results of the TFBS search with four models without taking into account site positions. It can be seen that all models jointly recognized 88.35 % of the peaks (3534 out of 4000, the sum of all areas within the Venn diagram, see Fig. 5, *a*, *b*). The overlap fraction of all models amounts 34.25 % (1370 out of 4000 peaks). Two non-PWM models BaMM and InMoDe make the significant contributions to peak recognition. They totally add 34.55 % of all peaks (696 + 647 + 39 = 1382 out of 4000). This fraction is almost the same as the overlap fraction of all models (1370). The BaMM model makes the largest independent contribution to recognition of sites, it adds 17.4 % of the peaks (696), in contrast to other models that add 0.525 % (21), 0.975 % (39) and 0.2 % (8) (PWM, InMoDe and diPWM respectively).

To assess the structural diversity of the TFBS, we build logos for peak fractions "only PWM", "only diPWM", "only BaMM", "only InMoDe" and "all models" (see Fig. 5, *c*). All logos contain the GTAAACA consensus. However, the "only PWM", "only diPWM" and "only InMoDe" fractions have the higher occurrence of GT than AT at the first two nucleotides of the consensus. It can also be noted that the 5′-ends of all logos are diverse in nucleotide content.

Classification of ChIP-seq peaks with taking into account the intersection of TFBS positions found by different models

The classification of peaks described above (without taking into account the positions of the TFBS) does not take into account positions of sites in peaks. To consider this circumstance we classify peaks with taking into account positions. We perform this for each pair of models (PWM–diPWM, PWM–BaMM, PWM–InMoDe, diPWM–BaMM, diPWM–InMoDe, InMoDe–BaMM). The results of the classification of peaks for GSE90454.FOXA2.KerCT are shown as the pie charts in Fig. 6.

All pairs of model combinations have very small fraction of "without intersection" peaks, ranging from 0.3 to 6.9 %. On the other hand, all cases were characterized by the large fraction of peaks "with intersection" (BaMM–InMoDe 53.6 %, PWM–diPWM 44.4 %, diPWM–BaMM 41.0 %, PWM–BaMM 37.3 %, diPWM–InMoDe 35.4 %, PWM–InMoDe 31.6 %). This fraction is larger for methodologically close pairs of models BaMM–InMoDe and PWM–diPWM (see Fig. 6). The fraction of the peaks with TFBS found with only a single model is the highest for BaMM model. In pairs of models PWM–BaMM, diPWM–BaMM, and InMoDe–BaMM, BaMM contributes greatly (39.2, 36.4 and 26.8 %, respectively).

Evaluation of the recognition TFBS accuracy by different models for FOXA2

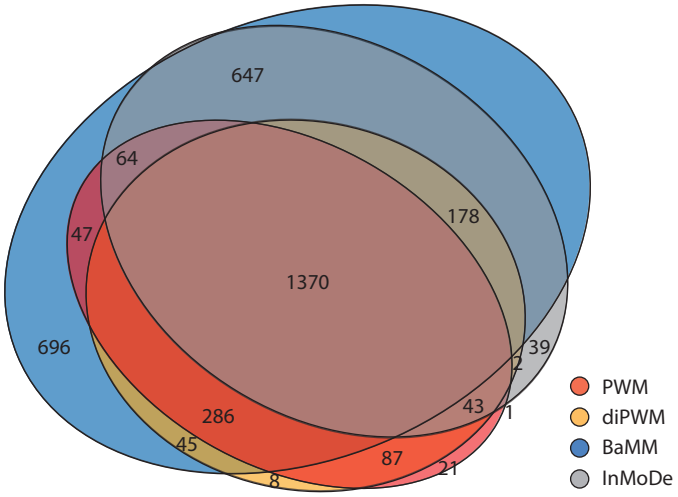
To compare the recognition accuracy of different models we calculate pAUC values from ROC curves obtained with the cross-validation procedure (see "Training *de novo* models and assessing the TFBS recognition accuracy") (Fig. 7, *a*). According to the results obtained, the values of the pAUC medians for the PWM, diPWM, BaMM and InMoDe models are 8.0E–4, 8.1E–4, 7.3E–4, and 5.6E–4, respectively. The differences between pAUC values were not significant ($p > 0.05$) for paired comparisons of PWM, diPWM, and BaMM, but the InMoDe model has significantly less values than any other model ($p < 0.05$).

Comparison of fractions of peaks with TFBS found by each model with that for all models. To investigate contributions of different models to the efficiency of TFBS search and to evaluate the overall result of several models, we determine fractions of peaks containing at least

a

Peaks with sites found by														
all models	all models				only two models						only one model			
	except PWM	except diPWM	except BaMM	except InMoDe	PWM diPWM	PWM BaMM	PWM InMoDe	diPWM BaMM	diPWM InMoDe	BaMM InMoDe	PWM	diPWM	BaMM	InMoDe
1370	178	64	43	286	87	47	1	45	2	647	21	8	696	39

b



c

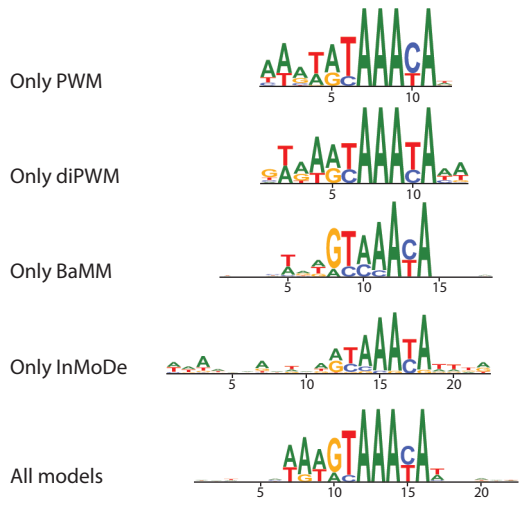


Fig. 5. The classification of peaks of the GSE90454.FOXA2.KerCT ChIP-seq dataset according to prediction results of all four models.
(*a*) Table, (*b*) Venn diagram, (*c*) Logo for fraction of the peaks respecting to predictions of sole models and that for the overlapping fraction of all models.

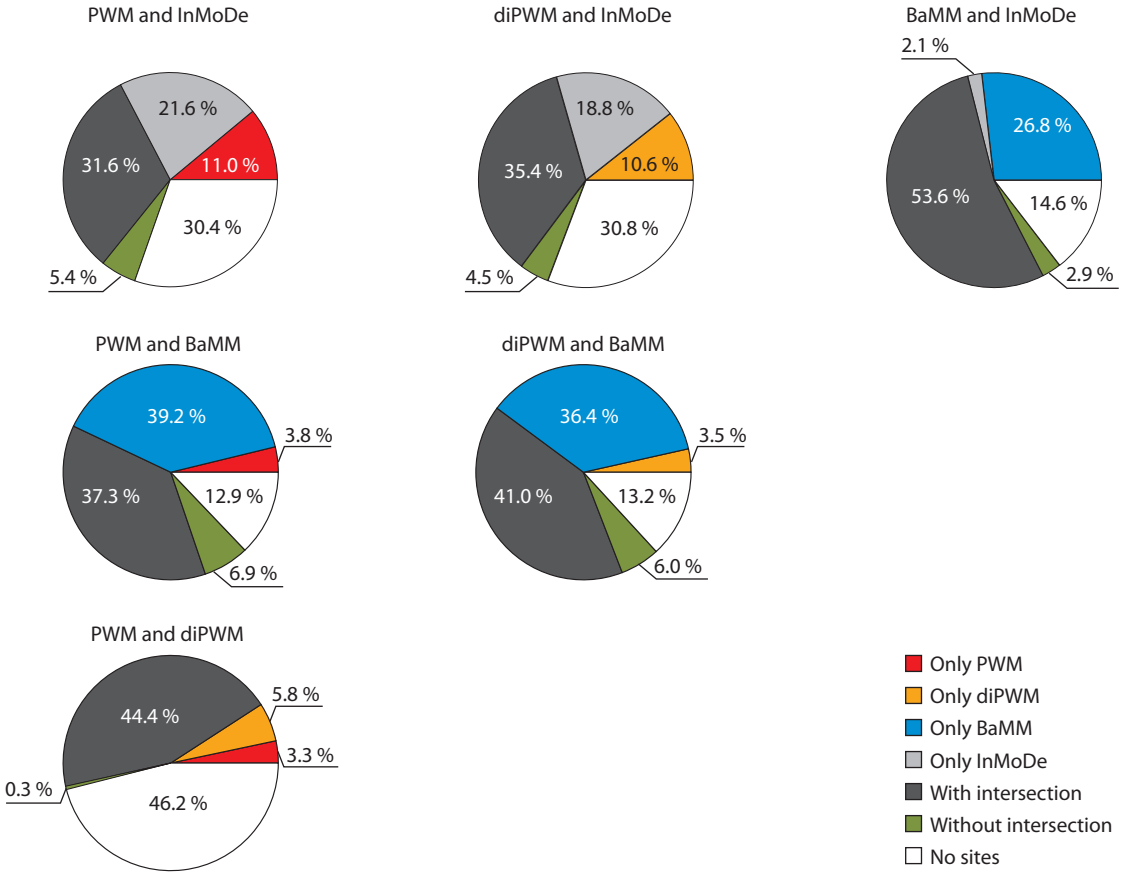


Fig. 6. Classification of the GSE90454.FOXA2.KerCT ChIP-seq dataset with taking into account intersection of TFBS positions respecting to different models.

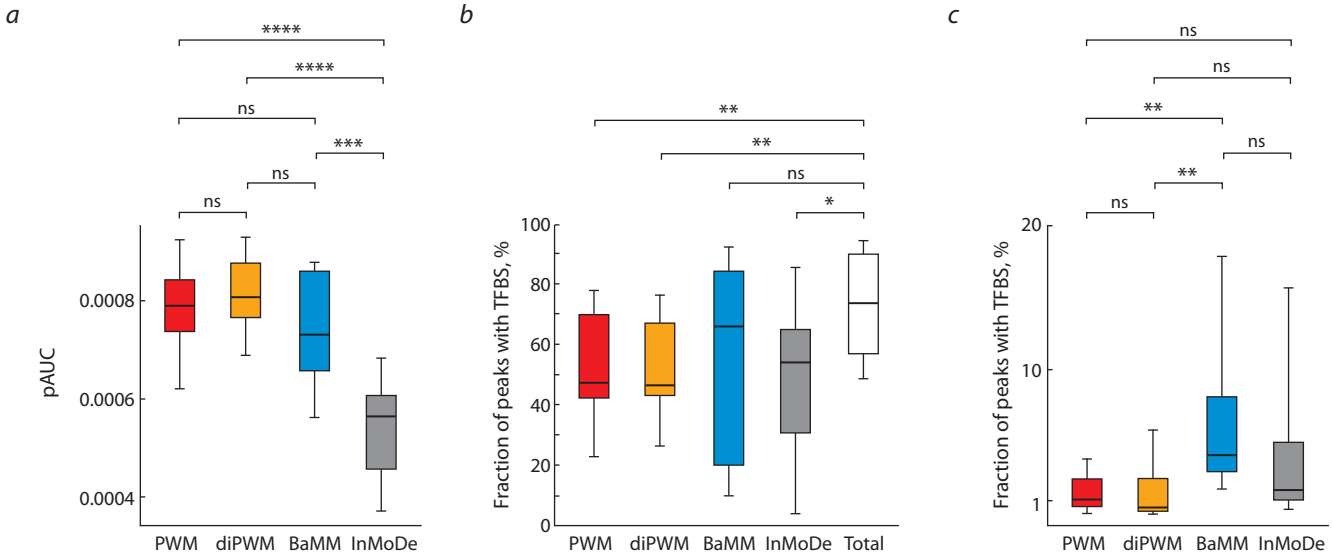


Fig. 7. The distribution of quartiles for recognition measures. The bottom part of the box denotes the minimum value of parameter; the top part denotes the maximum value of parameter. (*a*) Values of pAUC for all models. (*b*) Fractions of peaks recognized with a single models (PWM, diPWM, BaMM, InMoDe) and with all models together (Total). (*c*) Fractions of peaks contained only TFBS recognized with a single model.

ns – $p > 0.05$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

one TFBS for each sole model and those for all models together (see Fig. 7, *b*). The medians of recognized peaks fractions are 47.3, 46.4, 65.8, and 54 % for sole PWM, diPWM, BaMM and InMoDe, respectively. The median of recognized peaks fraction of joined results of all four models' case is 73.6 %. Consequently, together, all models add 26.3 % peaks containing TFBS to the fraction of sole PWM model, which is consistent with the earlier obtained result of using two fundamentally different PWM and SiteGA models (Levitsky et al., 2014). At the same time, the median values respecting fractions for the PWM, diPWM, and InMoDe models significantly lower ($p < 0.05$) than that obtained by combining all models. Thus, the approach using the combination of different models allows better identification of peaks with TFBS for FOXA2 than that using only one model. However, for BaMM, the fraction of recognized peaks did not statistically differ ($p > 0.05$) from the result obtained by combining the four models. Hence, the BaMM model makes the main contribution to the recognition of FOXA2 peaks and, among the other models this model better describes the structure of FOXA2 sites. However, the rest three models add 7.8 % of the peaks to the BaMM result, which proves the effectiveness of using different models together.

Comparison of fractions of peaks containing TFBS found by single models. As it is shown above, the combination of different models increases the number of peaks with TFBS. Hence, each model recognizes TFBS that others do not. To assess the independent contributions of all models to the search for TFBS, we determine the fractions of peaks containing TFBS of only one model (see Fig. 7, *c*). As one can see, each model (PWM, diPWM, BaMM, InMoDe) is able to find TFBS that other models do not find. The medians of peaks containing TFBS respecting a single model are 1.08, 0.49, 4.15, and 1.73 %, respectively for PWM, diPWM, BaMM, and InMoDe. At the same time, the results for BaMM are significantly different ($p < 0.05$) from those for both PWM and diPWM. It also confirms the assumption that the BaMM model better recognizes FOXA2 sites. However, each model contributes to site recognition. Consequently, each model reveals certain structural variant of TFBS.

Cross-validation test for PWM models with participation of their own training dataset and other ChIP-seq datasets

To estimate the dependence of specificity of various models for different ChIP-seq datasets as a function of a selection of particular dataset as the training sample, we performed the cross-validation test as follow. The accuracy of each PWM model we assessed not only within the same ChIP-seq training dataset, but also for the rest 15 datasets (control datasets). For the case of training dataset, we performed several iterations to divide the total training sample into 90 % of the peaks that we used to build a model, and the remaining 10 % of the peaks we used to estimate the performance. For each case we calculated the accuracy estimate pAUC (see "Training *de novo* models and assessing the TFBS recognition accuracy"), the results we presented in the form of the heatmap (Fig. 8). The heatmap shows that only in three cases ENCSR000BRE.A-549, ENCSR000BNI.Hep-G2 and ERP008682.pancreas other models have very low pAUC scores, and for five cases GSE90454.A-549, ENCSR066EBK.Hep-G2, GSE90454.KerCT, ENCSR080XEY.liver and ENCSR310NYI.liver, all models have high pAUC values.

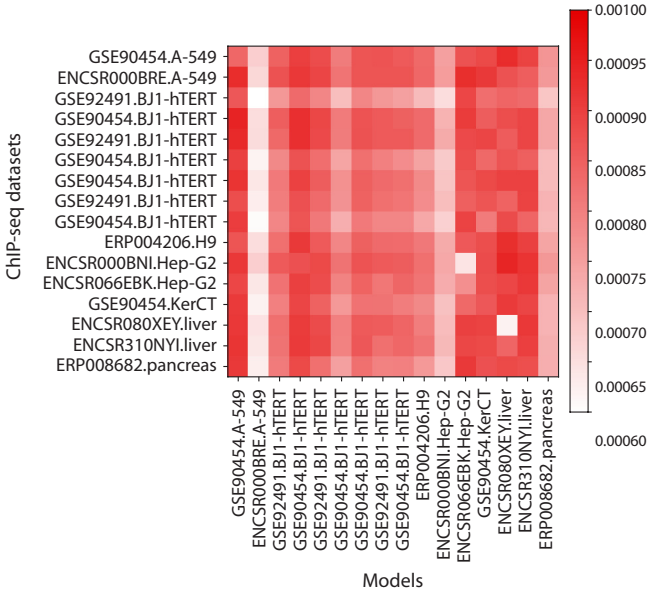


Fig. 8. The heatmap of cross-validation test results for PWM models.

Colors mark pAUC values. Each diagonal cell implies that control and training datasets are the same. Remaining cells refer to distinct training and control datasets. Rows mean models and columns denote ChIP-seq datasets.

Discussion

Based on all data obtained, we conclude that the joint use of alternative models allows us to expand the number of detected peaks containing TFBS relative to application of sole PWM.

This result can be explained by the presence of different structural types of TFBS of FOXA2. This is in agreement with experimental data obtained for a number of other TFs, including members of the FOX family. Thus, it was shown that HOXB13 and FOXC2 are able to bind with the same affinity to completely different sequences CAATAAA/TCGTAAA (Morgunova et al., 2018) and GTAAACA/ACAAATA (Chen et al., 2019), respectively. It was recently found that TF FOXN3 is able to bind to two fundamentally different types of TFBS, which had different lengths (Rogers et al., 2019). In addition, small changes in the structure of the TFBS depend on the cooperative interaction between TFs (Morgunova, Taipale, 2017). Hence, we propose that FOXA2 also can bind to different structural types of BS.

To take into account all the TFBS structural types, the only PWM model for site recognition may not be enough, this problem partially was solved using several PWMs (Bi et al., 2011; Mitra et al., 2018) or using alternative models (Mathelier, Wasserman, 2013; Yang et al., 2014; Siebert, Söding, 2016; Eggeling et al., 2017; Gheorghe et al., 2019). However, previously alternative models were usually compared with PWM only in terms of the recognition accuracy (Siebert, Söding, 2016), or according the number of recognized TFBSs (Samee et al., 2019). In the current study, we took in analysis FOXA2 ChIP-seq data. We compared not only the accuracy and the number of peaks recognized, but also we estimated independent contributions of each model and assessed the joint contribution for all pairs of models, and also we tested positions of hits in peaks for each pair of models. The results for the accuracy assessment (see Fig. 7, *a*) showed that the InMoDe model had the lowest accuracy relative to other models, and the BaMM, diPWM and PWM models were comparable in accuracy. In terms of expanding the total fraction of peaks with TFBS, the BaMM model performed the best, since this model found the largest fraction of peaks with TFBS that other models do not find. Nevertheless, all alternative diPWM, BaMM and InMoDe models allow expanding the pool of recognized TFBS relative to sole PWM, but PWM also makes an independent contribution to the total number of peaks with recognized TFBS.

Conclusion

We have developed the pipeline MultiDeNA, which allows uniform processing of ChIP-seq data using different TFBS models. Currently, it can be used to build PWM, diPWM, InMoDe, BaMM models. MultiDeNA includes the steps of preparing data, building models, evaluating recognition accuracy, scanning peaks, combining results, and analyzing them. The developed pipeline of programs processed datasets from the ReMap database, including 22 ChIP-seq experiments for TF FOXA2. We have shown that combined use of different models allows increasing the total fraction of recognized peaks up to 73.6 % (relative to sole PWM model, the fraction of recognized peaks increased by 26.3 %). We have shown that different models tend to recognize the same sites of FOXA2 in the large fraction of peaks, thereby revealing the most common structural type of TFBS in these peaks. Also, each model found TFBS that other models did not predict. The BaMM model performed the best with 4.15 % of peaks containing only its sites, versus 1.08, 0.49, 1.73 % for PWM, diPWM and InMoDe, respectively. We proposed that the heterogeneity of sites for FOXA2 is revealed only if two or more models are applied. The diPWM model showed worst result in sole application in comparison with other models (diPWM recognized TFBS in 46.4 % of the peaks). The best model for the FOXA2 sites was BaMM; it found TFBS in 65.8 % of the peaks. Hence, we assumed that the BaMM model could better describe BS for FOXA2.

References

Bailey T.L., Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proc. Int. Conf. Intell. Syst. Mol. Biol. 1994;2:28-36. DOI citeulike-article-id:878292. PMID 7584402.

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(20):4442-4451. DOI 10.1093/nar/gkf578.

Bi Y., Kim H., Gupta R., Davuluri R.V. Tree-based position weight matrix approach to model transcription factor binding site profiles. *PLoS One.* 2011;6(9):e24210. DOI 10.1371/journal.pone.0024210.

Bulyk M.L., Johnson P.L.F., Church G.M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* 2002;30(5):1255-1261. DOI 10.1093/nar/30.5.1255.

Chen X., Wei H., Li J., Liang X., Dai S., Jiang L., Guo M., Qu L., Chen Z., Chen L., Chen Y. Structural basis for DNA recognition by FOXC2. *Nucleic Acids Res.* 2019;47(7):3752-3764. DOI 10.1093/nar/gkz077.

Chèneby J., Ménétrier Z., Mestdagh M., Rosnet T., Douida A., Rhalloussi W., Bergon A., Lopez F., Ballester B. ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.* 2020;48(D1):D180-D188. DOI 10.1093/nar/gkz945.

Eggeling R., Grosse I., Grau J. InMoDe: tools for learning and visualizing intra-motif dependencies of DNA binding sites. *Bioinformatics.* 2017;33(4):580-582. DOI 10.1093/bioinformatics/btw689.

Farnham P.J. Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* 2009;10(9):605-616. DOI 10.1038/nrg2636.

Furey T.S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* 2012;13(12):840-852. DOI 10.1038/nrg3306.

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(4):e21. DOI 10.1093/nar/gky1210.

Gupta S., Stamatoyannopoulos J.A., Bailey T.L., Noble W.S. Quantifying similarity between motifs. *Genome Biol.* 2007;8(2):R24. DOI 10.1186/gb-2007-8-2-r24.

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(4):576-589. DOI 10.1016/j.molcel.2010.05.004.

Ignatieva E.V., Oshchepkov D.Y., Levitsky V.G., Vasiliev G.V., Klimova N.V., Busygina T.V., Merkulova T.I. Comparison of the results of search for the SF-1 binding sites in the promoter regions of the steroidogenic genes, using the SiteGA and SITECON methods. In: Proc. Fourth Int. Conf. Bioinform. Genome Regul. Struct. (BGRS). 2004;1:69-72.

Iwafuchi-Doi M. The mechanistic basis for chromatin regulation by pioneer transcription factors. *WIREs Syst. Biol. Med.* 2019;11(1):e1427. DOI 10.1002/wsbm.1427.

Keilwagen J., Grau J. Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.* 2015;43(18):e119. DOI 10.1093/nar/gkv577.

Kiesel A., Roth C., Ge W., Wess M., Meier M., Söding J. The BaMM web server for de-novo motif discovery and regulatory sequence analysis. *Nucleic Acids Res.* 2018;46(W1):W215-W220. DOI 10.1093/nar/gky431.

Kulakovskiy I.V., Boeva V.A., Favorov A.V., Makeev V.J. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics.* 2010;26(20):2622-2623. DOI 10.1093/bioinformatics/btq488.

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(01):1340004. DOI 10.1142/S0219720013400040.

Kulakovskiy I.V., Makeev V.J. Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources. *Biophysics* (*Oxf.*). 2009;54(6):667-674. DOI 10.1134/S0006350909060013.

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., Kolpakov F.A., Makeev V.J. 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(D1):D252-D259. DOI 10.1093/nar/gkx1106.

Lambert S.A., Jolma A., Campitelli L.F., Das P.K., Yin Y., Albu M., Chen X., Taipale J., Hughes T.R., Weirauch M.T. The human transcription factors. *Cell.* 2018;172(4):650-665. DOI 10.1016/j.cell.2018.01.029.

Latchman D.S. Transcription factors: bound to activate or repress. *Trends Biochem. Sci.* 2001;26(4):211-213. DOI 10.1016/S0968-0004(01)01812-6.

Levitsky V.G., Ignatieva E.V., Ananko E.A., Turnaev I.I., Merkulova T.I., Kolchanov N.A., Hodgman T.C.T. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. *BMC Bioinform.* 2007;8(1):1-20. DOI 10.1186/1471-2105-8-481.

Levitsky V.G., Kulakovskiy I.V., Ershov N.I., Oshchepkov D.Y., Makeev V.J., Hodgman T.C., Merkulova T.I. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genom.* 2014;15(1):80. DOI 10.1186/1471-2164-15-80.

Levitsky V.G., Oshchepkov D.Y., Klimova N.V., Ignatieva E.V., Vasiliev G.V., Merkulov V.M., Merkulova T.I. Hidden heterogeneity of transcription factor binding sites: a case study of SF-1. *Comput. Biol. Chem.* 2016;64:19-32. DOI 10.1016/j.compbiolchem.2016.04.008.

Lloyd S.M., Bao X. Pinpointing the genomic localizations of chromatin-associated proteins: the yesterday, today, and tomorrow of ChIP-seq. *Curr. Protoc. Cell Biol.* 2019;84(1):e89. DOI 10.1002/cpcb.89.

Machanick P., Bailey T.L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics.* 2011;27(12):1696-1697. DOI 10.1093/bioinformatics/btr189.

Mathelier A., Wasserman W.W. The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.* 2013;9(9):e1003214. DOI 10.1371/journal.pcbi.1003214.

McClish D.K. Analyzing a portion of the ROC curve. *Med. Decis. Mak.* 1989;9(3):190-195. DOI 10.1177/0272989X8900900307.

Mitra S., Biswas A., Narlikar L. DIVERSITY in binding, regulation, and evolution revealed from high-throughput ChIP. *PLoS Comput. Biol.* 2018;14(4):1-20. DOI 10.1371/journal.pcbi.1006090.

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

Morgunova E., Yin Y., Das P.K., Jolma A., Zhu F., Popov A., Xu Y., Nilsson L., Taipale J. Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. *eLife.* 2018;7:1-21. DOI 10.7554/eLife.32963.

Park P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 2009;10(10):669-680. DOI 10.1038/nrg2641.

Quinlan A.R., Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841-842. DOI 10.1093/bioinformatics/btq033.

Rogers J.M., Waters C.T., Seegar T.C.M., Jarrett S.M., Hallworth A.N., Blacklow S.C., Bulyk M.L. Bispecific forkhead transcription factor FoxN3 recognizes two distinct motifs with different DNA shapes. *Mol. Cell.* 2019;74(2):245-253.DOI 10.1016/j.molcel. 2019.01.019.

Samee M.A.H., Bruneau B.G., Pollard K.S. A *de novo* shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Syst.* 2019;8(1):27-42. DOI 10.1016/j.cels.2018.12.001.

Siebert M., Söding J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.* 2016;44(13):6055-6069. DOI 10.1093/nar/gkw521.

Srivastava D., Mahony S. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochim. Biophys. Acta – Gene Regul. Mech.* 2020;1863(6):e194443. DOI 10.1016/j.bbagrm.2019.194443.

Stormo G.D. DNA binding sites: representation and discovery. *Bioinformatics.* 2000;16(1):16-23. DOI 10.1093/bioinformatics/16.1.16.

Wallerman O., Motallebipour M., Enroth S., Patra K., Bysani M.S.R., Komorowski J., Wadelius C. Molecular interactions between

HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Res.* 2009;37(22):7498-7508. DOI 10.1093/nar/gkp823.

Wederell E.D., Bilenky M., Cullum R., Thiessen N., Dagpinar M., Delaney A., Varhol R., Zhao Y., Zeng T., Bernier B., Ingham M., Hirst M., Robertson G., Marra M.A., Jones S., Hoodless P.A. Global analysis of *in vivo* Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.* 2008;36(14):4549-4564. DOI 10.1093/nar/gkn382.

Worsley Hunt R., Wasserman W.W. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.* 2014;15(7):412. DOI 10.1186/s13059-014-0412-4.

Yang L., Zhou T., Dror I., Mathelier A., Wasserman W.W., Gordân R., Rohs R. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.* 2014;42(D1):D148-D155. DOI 10.1093/nar/gkt1087.

Zhang M.O., Marr T.G. A weight array method for splicing signal analysis. *Bioinformatics.* 1993;9(5):499-509. DOI 10.1093/bioinformatics/9.5.499.

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nusbaum C., Myers R.M., Brown M., Li W., Liu X.S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137. DOI 10.1186/gb-2008-9-9-r137.

ORCID ID

A.V. Tsukanov orcid.org/0000-0002-5174-6609
V.G. Levitsky orcid.org/0000-0002-4905-3088

Acknowledgements. This work was supported by the Russian Foundation for Basic Research No. 18-29-13040 and the state budget project No. 0259-2019-0008.

Conflict of interest. The authors declare no conflict of interest.

Received October 10, 2020. Revised January 10, 2021. Accepted January 12, 2021.