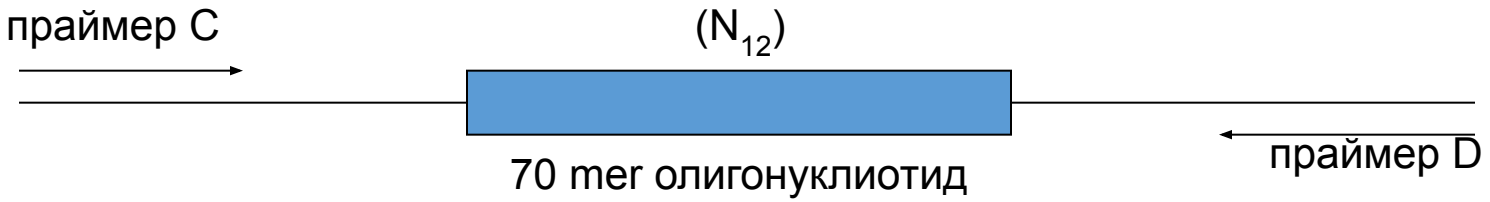
Методы изучения регуляторных районов генов

Лекция III

Меркулова Татьяна Ивановна

Институт цитологии и генетики СО РАН



70 mer олигонуклиотид

70 вторая синтезированная цепь

Мечение двойной цепи

Набор меченых 70bp олигонуклеотидов пропускается через аффинную колонку, содержащую Hoxd1 белок

Эллюция связанных олигонуклиотидов

Выделение эллюированных олигонуклиотидов

ПЦР-амплификация с выделенных олигонуклиотидов

3 цикла

Клонирование и секвенирование отобранных олигонуклиотидов

Method

Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities

Arttu Jolma,[1,2] Teemu Kivioja,[1,3] Jarkko Toivonen,[3] Lu Cheng,[3] Gonghong Wei,[1] Martin Enge,[2] Mikko Taipale,[1] Juan M. Vaquerizas,[4] Jian Yan,[1] Mikko J. Sillanpää,[5] Martin Bonke,[1] Kimmo Palin,[3] Shaheynoor Talukder,[6] Timothy R. Hughes,[6] Nicholas M. Luscombe,[4] Esko Ukkonen,[3] and Jussi Taipale[1,2,7]

[1]Department of Molecular Medicine, National Public Health Institute (KTL) and Genome-Scale Biology Program, Institute of Biomedicine and High Throughput Center, University of Helsinki, Biomedicum, Helsinki, Finland; [2]Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden; [3]Department of Computer Science, FI-00014 University of Helsinki, Helsinki, Finland; [4]EMBL–European Bioinformatics Institute, Cambridge CB10 1SD, United Kingdom; [5]Department of Mathematics and Statistics, FI-00014 University of Helsinki, Helsinki, Finland; [6]Department of Molecular Genetics and Banting and Best Department of Medical Research, University of Toronto, Toronto, ON M4T 2J4, Canada

Genome Research
www.genome.org

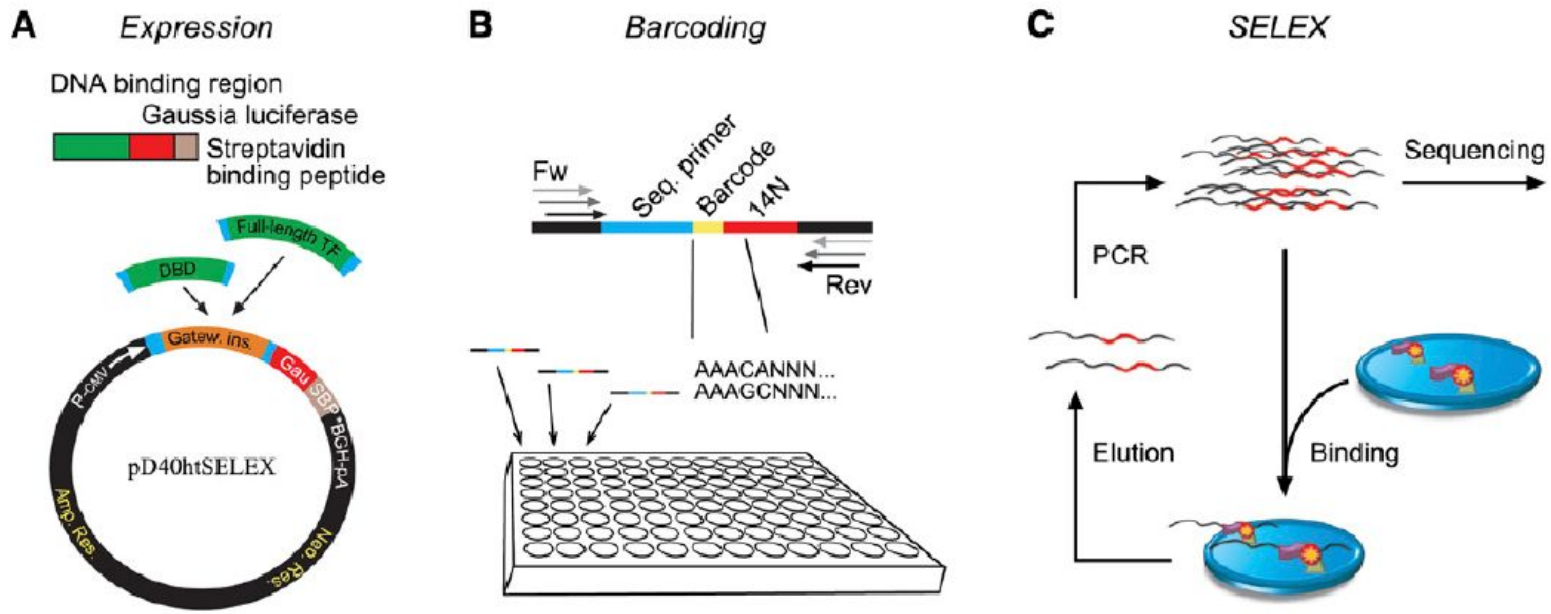


Figure 1. Schematic description of the high-throughput SELEX process. (*A*) Protein expression. (*Top*) Proteins are expressed as fusion proteins with SBP-tagged *Gaussia*-luciferase. (*Bottom*) The GATEWAY recombination cloning system is used to transfer DNA sequences encoding DBDs or TFs from donor-vectors to the pD40htSELEX expression vector. (*B*) Ligand design that accommodates multiplexing of samples using barcodes. Each DNA ligand contains a 14-bp randomized region (14N), and a 5-bp barcode (Barcode) that uniquely identifies the individual SELEX sample. To increase specificity, each barcode differs from all other barcodes by at least 2 bp. These variable sequences are flanked by constant sequences that include an Illumina Genome Analyzer sequencing primer site (Seq. primer) and bridge amplification primer binding regions (Fw, Rev; arrows), which are extended in their 5′ regions to accommodate partially nested primers (used in successive SELEX rounds). (*C*) Basic principle of high-throughput SELEX. A double-stranded DNA mixture containing all possible 14-bp sequences (from *B*) is incubated with a DNA-binding protein immobilized into a well of a 96-well plate, resulting in binding of DNA to the protein. After washing and elution, the resulting population of more specific sequences is amplified by PCR and subjected to high-throughput single-molecule sequencing. The specificity of the TF can then be constructed by iterating the process and calculating the abundance of distinct sequences after different numbers of cycles. In each cycle, multiple reactions are mixed into a single sequencing lane, and the TFs are identified using the barcode sequences.

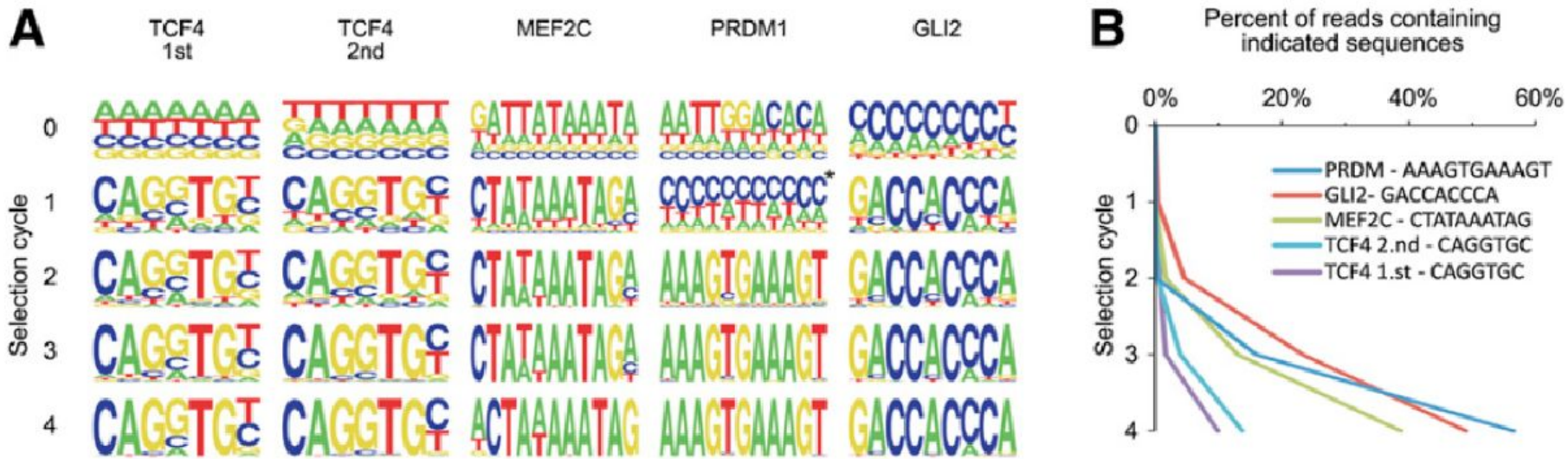


Figure 2. Enrichment of specific sequences during the SELEX process. (*A*) Position weight matrices built around the most enriched sequence for four different TFs (see Methods for details). The height of the letter at each position is directly proportional to the incidence of the indicated base in sequences where all other bases exactly match the most enriched sequence. Note that clear enrichment of sequences is observed after one or two SELEX rounds, and that two separate experiments for TCF4 result in a very similar enrichment pattern. In the first cycle, the algorithm used here detects incorrect binding profile for PRDM1 (asterisk) due to a low number of the relatively long consensus sequences. The enrichment of high-affinity sequences can, however, be detected by seeding the algorithm with consensus from the later cycles (see Supplemental Fig. S4A). (*B*) The fraction of all fragments containing the most enriched sequence from the third SELEX cycle plotted as a function of the SELEX cycle.

A E-Box (CANNTG)

```
A1                gcgtg CTGGTGT CACGTG TTT ggaga
A2                   gcgtg TAT CACGTG AGCGTTC ggaga
A3                 gcgtg GGTCAT CACGTG AGCA ggaga
A4              tctcc ATATAGTAT CACGTG A cacgc
A5                   gcgtg TGAT CATGTG ATGCCC ggaga
A6                   tctcc ACAT CATGTG ATTATG cacgc
A7                   gcgtg GTAT CACCTG ACTAGC ggaga
A8                tctcc CAACGGAT CACCTG AT cacgc
A9                   tctcc TGAT CACATG CCAGTA cacgc
A10             gcgtg TTGATTCAT CACGTG C ggaga
A11             tctcc TGGCACAAT CACGTG A cacgc
A12               tctcc TTATAGT CACGTG ACT cacgc
A13               tctcc ATCTAGT CACGTG ACA cacgc
A14                 tctcc GCACT CACCTG ATTTG cacgc
A15             tctcc ACAGACGGT CACGTG A cacgc
A16               tctcc AACACAC CACGTG ACC cacgc
A17                 gcgtg GCGAT CACGTG TATTA ggaga
A18             tctcc ACAGATGAT CACGTG A cacgc
A19                 gcgtg TTCAT CACCTG AGTCA ggaga
A20                 gcgtg GCGAT CAGGTG ATGCA ggaga
A21                   gcgtg GGT CACGTG GTGTCCA ggaga
A22             tctcc CAAGCGGAT CACGTG A cacgc
A23             tctcc TGCACACTA CACGTG A cacgc
A24                   tctcc TGAT CACATG CCATGT cacgc
A26                gcgtg GGTCAT CACGTG AGCC ggaga
A27              gcgtg CAACGGAT CACCTG AT ggaga
A28             tctcc TCTGTTCAT CACGTG A cacgc
A29              tctcc TCATGCAT CACGTG AT cacgc
A30              gcgtg TGCACAAT CACGTG AT ggaga
```

Position	-6	-5	-4	-3	-2	-1	1	2	3	4	5
G	11	5				1	22		30	1	5
A	4	23	1		30		2			21	2
T	5	1	28			2	.	30		2	11
C	10	1	1	30		27	6			3	12
	G/C	A	T	C	A	C	G	T	G	A	Py

ADD1/SREBP1 E-Box Consensus Binding Sequence:

5'-ATCACGTGA-3'

B Non-E-Box

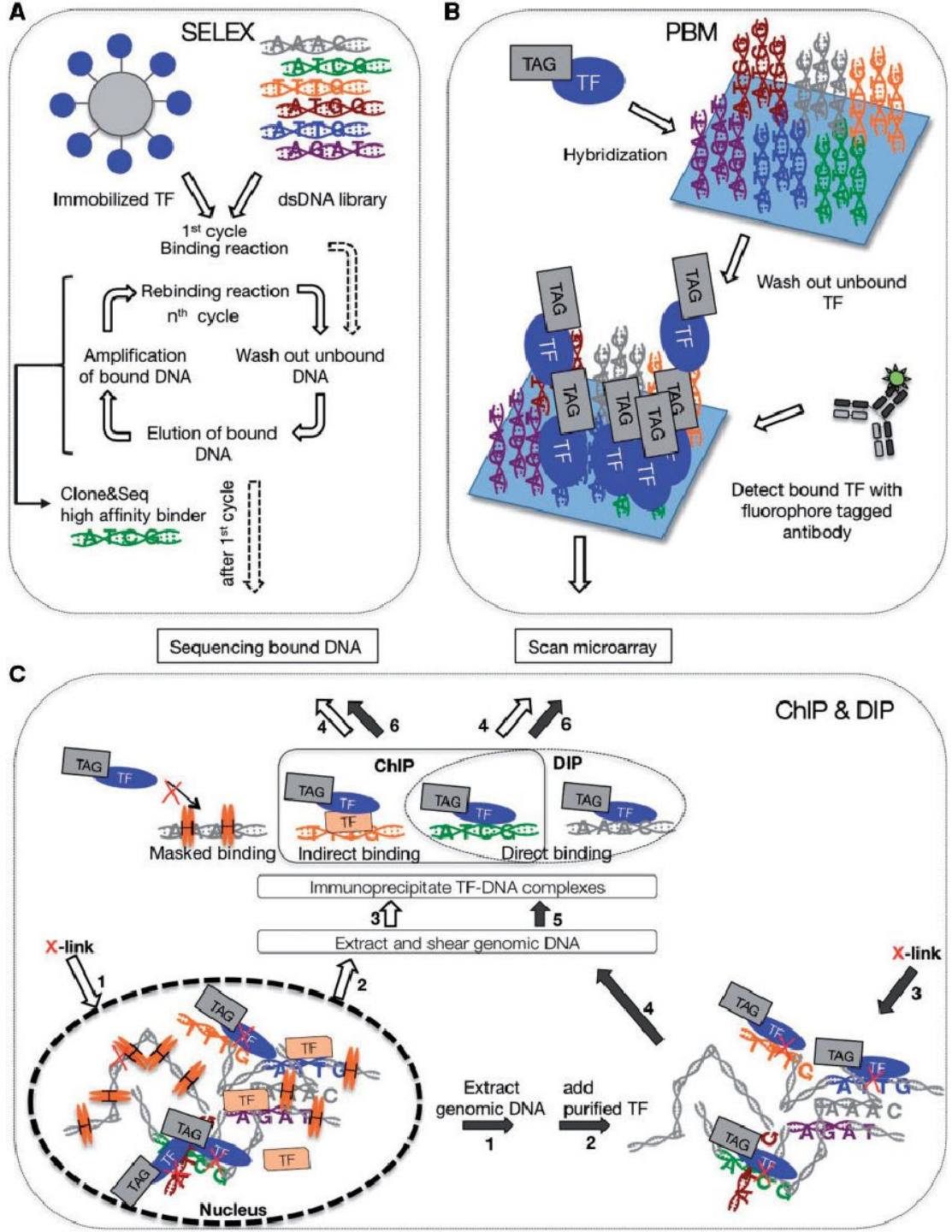
```
S1              tctcc GATCTTGAG ATCACCCcac gc
S2              tctcc AGAGCTAGT GTCACCCcac gc
S3              tctcc TTGTGTCTG ATCACCCcac gc
S4                    tctcc ATCACCCCAC CTGTCGcacgc
S5                gcgtg TCAGAG ATCACCCCAC ggaga
S6              gcgtg CTAGCTCGT GTCACCCcac gc
S7              tctcc ACGTATTGT ATCACCCcac gc
```

Position	-1	1	2	3	4	5	6	7	8	9	10
G	3	2									
A		5			7					7	
T	3		7								
C	1			7		7	7	7	7		7
	N	A	T	C	A	C	C	C	C	A	C

ADD1/SREBP1 Non-E-Box Consensus Binding Seque

SAAB Consensus sequence: 5'-ATCACCCCAC-3

Published SRE-1 sequence: 5'-ATCACCCCAC-3


A
SELEX
Immobilized TF
dsDNA library
1st cycle
Binding reaction
Rebinding reaction
nth cycle
Amplification
of bound DNA
Wash out unbound
DNA
Elution of bound
DNA
Clone&Seq
high affinity binder
after 1st cycle
Sequencing bound DNA
B
PBM
TAG
TF
Hybridization
Wash out unbound
TF
Detect bound TF with
fluorophore tagged
antibody
Scan microarray
C
ChIP & DIP
ChIP
DIP
Masked binding
Indirect binding
Direct binding
Immunoprecipitate TF-DNA complexes
Extract and shear genomic DNA
X-link
Nucleus
Extract
genomic DNA
add
purified TF

A

МТдт 5'-AGCACTATA**GG**G**ACA**TGA**TGTTCC**ACACGTCACATGGTCGTCC-3'

МТ27 5'-GATGTCGCG**GG**A**ACA**CAG**TGTTCC**GTGTACTGTGCAACTACTT-3'

25 5'-GATCCCCCCGGGCATCACCGTGCAGGGGGGGA **GGTACA**GAG**TGTTCT**
GCGAGGATGCG-3' G/C богатое окружение (28% A/T п.н.)

111 5'-GATCCTGTTACCATAGTGTAACTTCCTATTCA **GGTACA**ATA**TGTTCT**
ATACTTCTATTT-3' A/T богатое окружение (72% A/T п.н.)

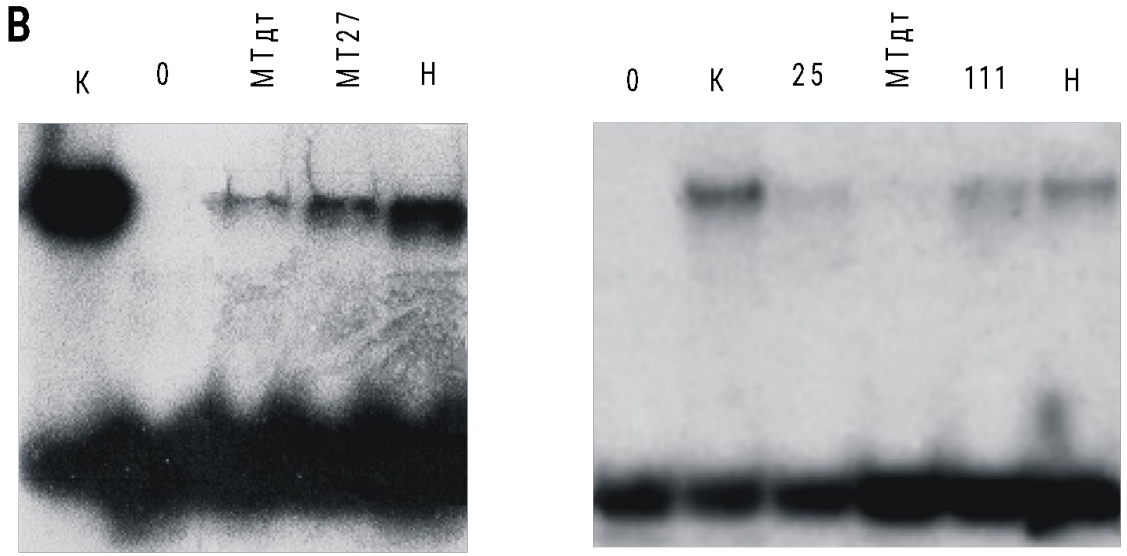
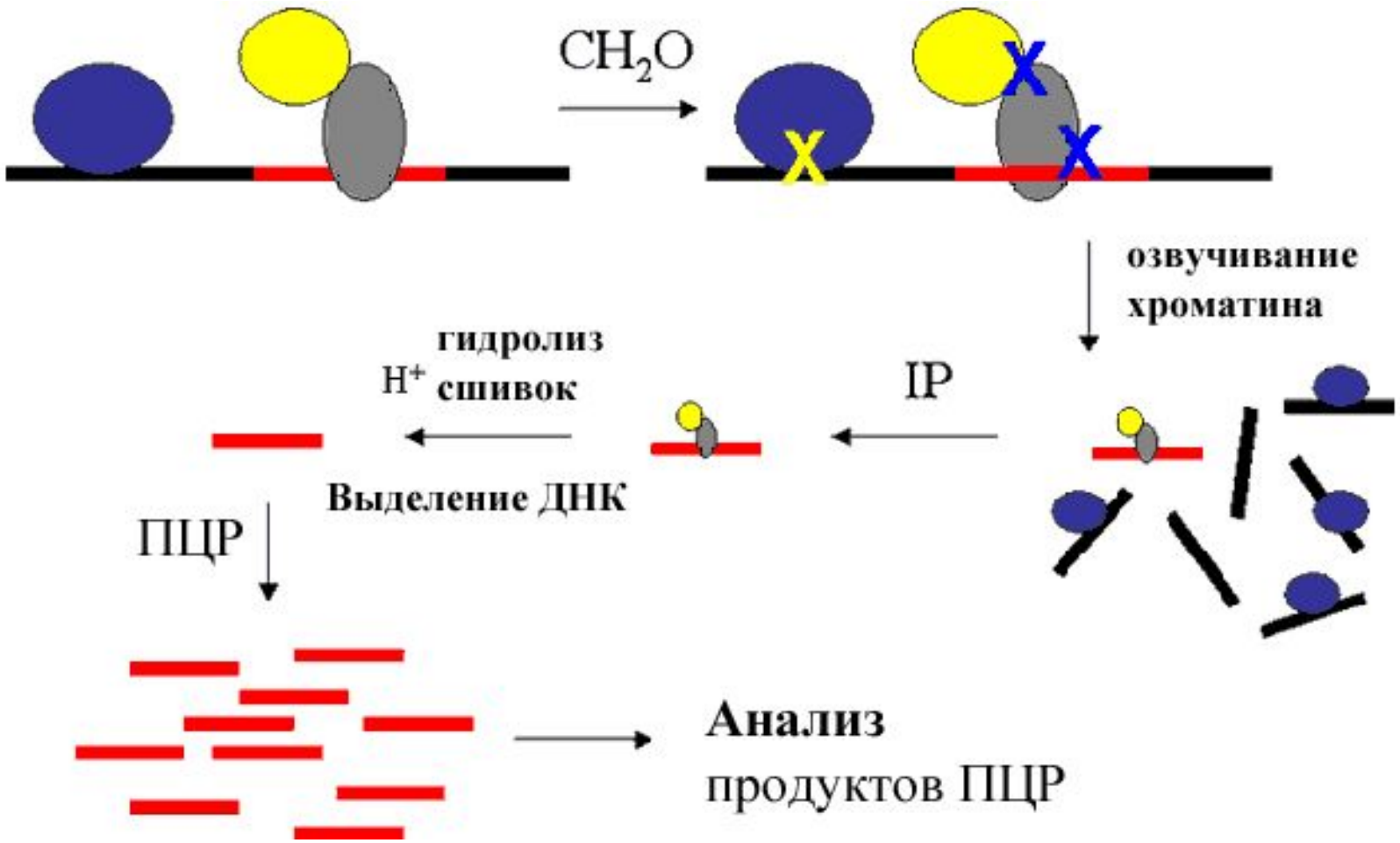


Рис. 2. Влияние первичной структуры ДНК в окружении консенсуса сайтов связывания рецептора глюкокортикоидов на эффективность связывания ДНК с рецептором.

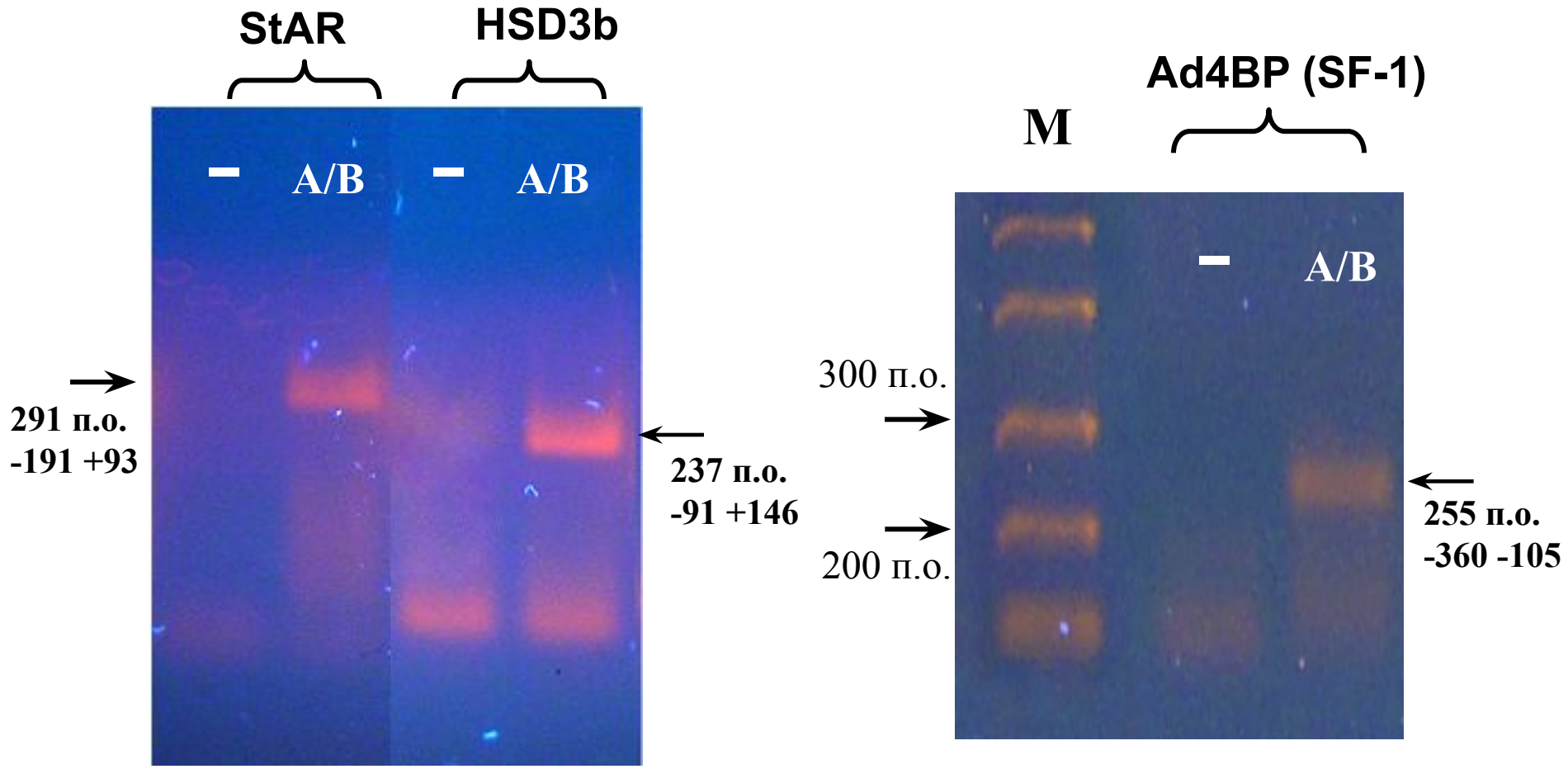
К - ДНК-проба в комплексе с рецептором глюкокортикоидов; 0 - свободная ДНК-проба; 3-5 и 8-11 - вытеснение меченой пробы из коплекса с ГР различными фрагментами немеченой конкурентной ДНК: МТдт - *Eco*RI-*Hind*III фрагмент pМТдт, МТ27 - *Eco*RI-*Hind*III фрагмент pМТ27, 25 - *Eco*RI-*Hind*III фрагмент p25, 111 - *Eco*RI-*Hind*III фрагмент p111, Н - *Pvu*II-*Eco*RI фрагмент pUC19 (неспецифическая ДНК).

Принцип метода иммунопреципитации *in vivo*

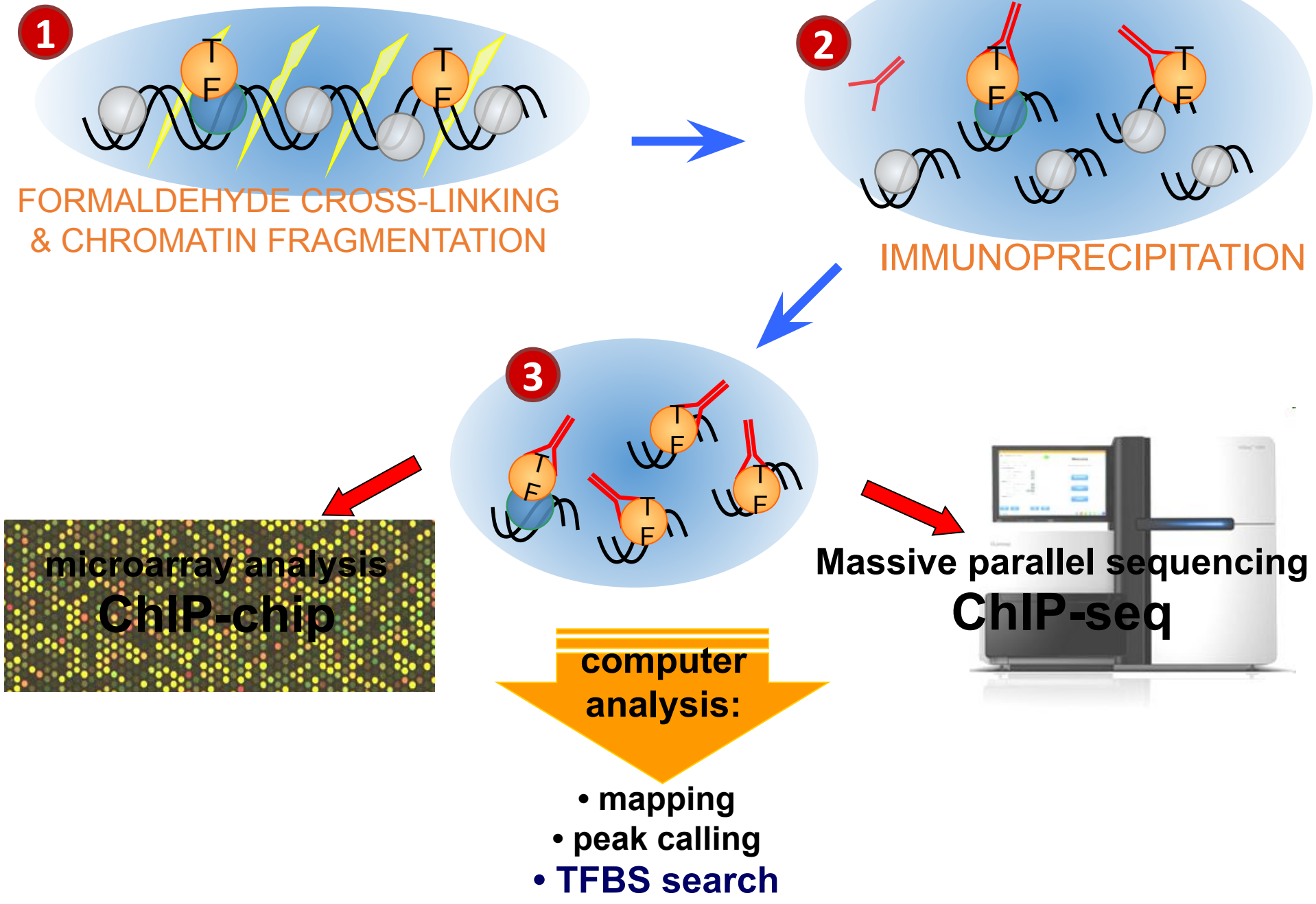


Иммунопреципитация хроматина *in vivo* с использованием антител против SF-1 для трёх генов мыши

Семенники мышей линии C57Br возраста 7-10 дней



Scheme of chromatin immunoprecipitation workflow



Features of p65-binding sites on chromosome 22

(*A*) p65 binding relative to Sanger-annotated genes and hybridizing regions on chromosome 22 is illustrated.

(*B*) Distribution of NF-B consensus sequences in p65-binding sites.

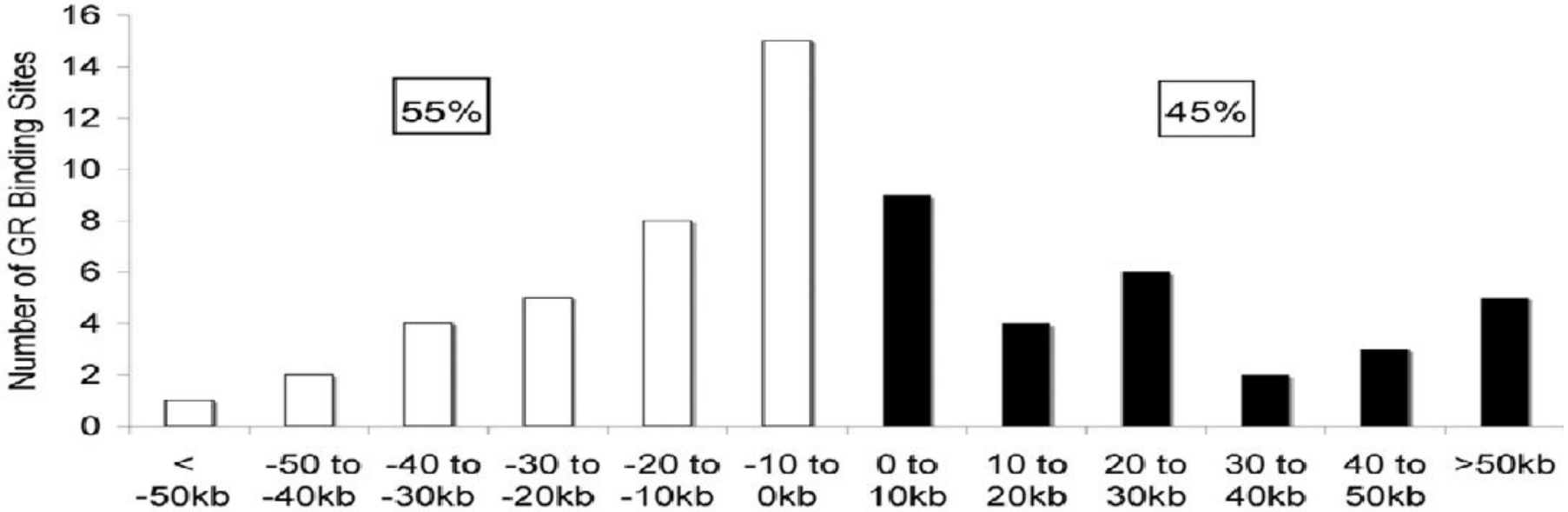
The sequences of p65-bound fragments on the microarray were searched for NF-B consensus sites by using both an in-house chromosome annotation system and the TFSEARCH database.

www.cbrc.jpresearchdbTFSEARCH.html

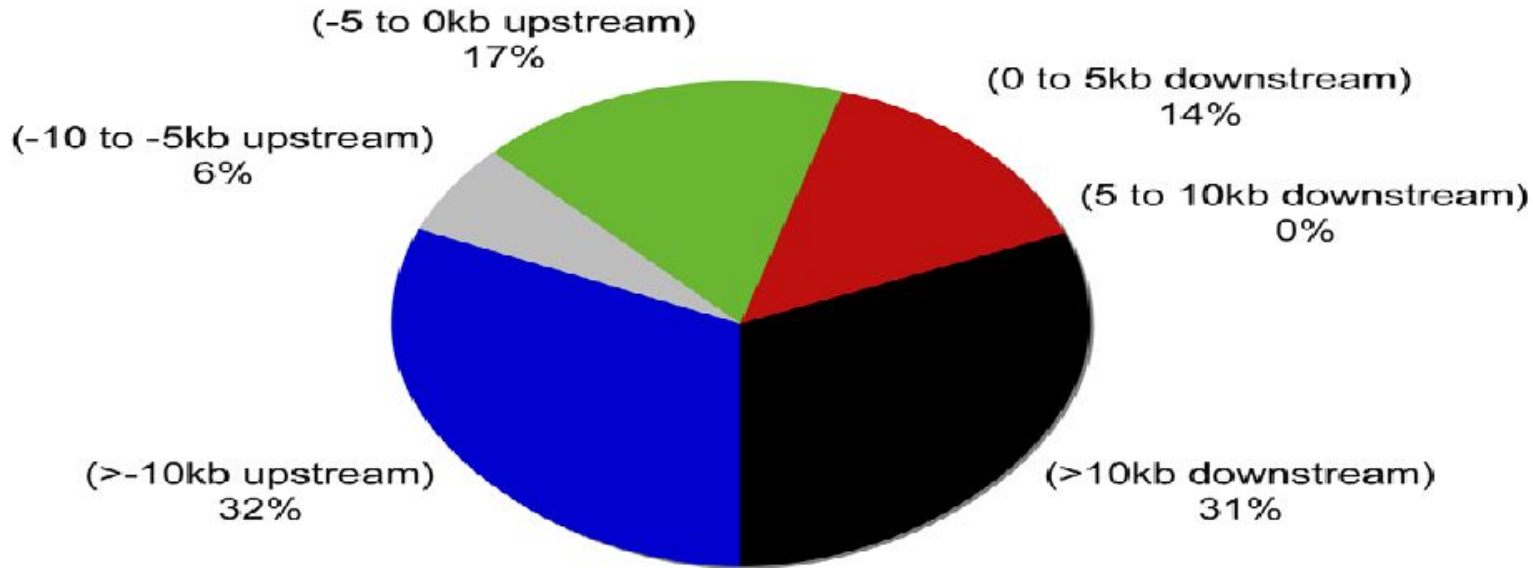
R. Martone, G. Euskirchen, P. Bertone, S. Hartman et al. **Distribution of NF-B-binding sites across human chromosome 22**
PNAS, October 14, 2003, vol. 100 no. 21, 12247–12252

Location and Position of GREs

A



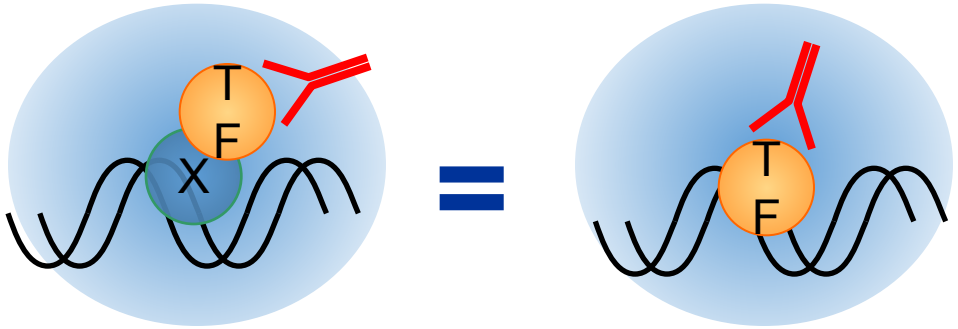
B



Weak points of TFBS identification by ChIP-seq technology

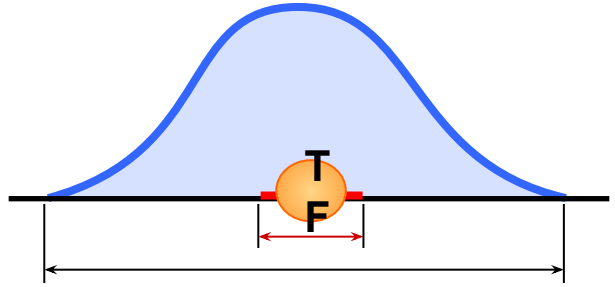
1 Indirect TF binding to DNA

The cases of direct and indirect interaction of TF with DNA are indistinguishable



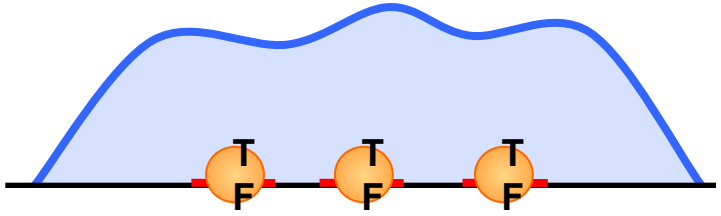
2 Insufficient resolution

Typical ChIP-seq peak – 100-500 bp
Typical TF footprint on DNA – 5-25 bp



3 TFBS clusters

A group of neighboring TFBSs is usually processed by peak-callers as one peak



4 False positive peaks (antibody cross-specifity, overrepresented seqs)

5 The spectrum of detected binding loci depends on cell type and state

Используемые для интерпретации данных ChIP-seq биоинформационные методы

Методы предсказания ССТФ

De novo методы поиска ССТФ

- **методы поиска наиболее представленных мотивов в участках связывания ТФ с хроматином (пиках ChIP-seq).**

(без предварительной информации о ССТФ)

monoCHiPMunk

Kulakovskiy I.V., et al., Bioinformatics 2010, 26(20):2622-3

diCHiPMunk

Методы, основанные на использовании обучающих выборок ССТФ

- Консенсус, весовые матрицы (PWM) и пр.

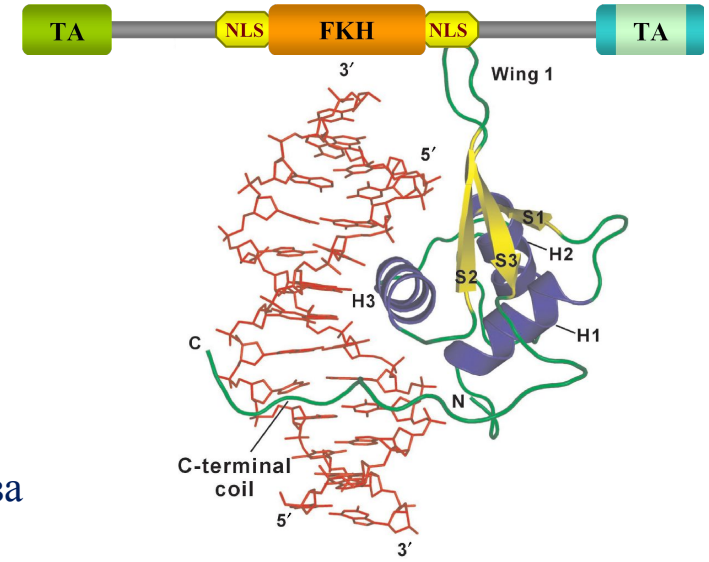
FoxA oPWM
SiteGA

[Levitsky V.G., et al., BMC Bioinformatics 2007, 8:481]
[Levitsky V.G., et al., Dokl Biochem Biophys. 2011, 436:12-5]

Объект исследования: сайты связывания транскрипционных факторов FoxA

ТФ FoxA : FoxA1, FoxA2, FoxA3

- FoxAs имеют высококонсервативный ДСД типа winged-helix (FKH).
- FoxA обладают способностью вытеснять линкерные гистоны и открывать компактизованный хроматин, т.е. могут служить как **pioneer transcription factors.**
- FoxAs играют важную роль в раннем эмбриогенезе, органогенезе и поддержании метаболического гомеостаза организма.
- Все три ТФ FoxA остаются активными в печени взрослого, играя активную роль в регуляции метаболизма

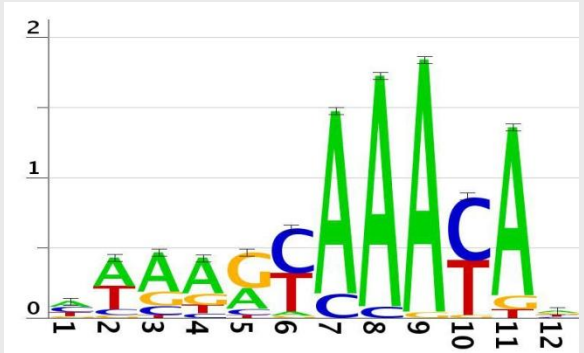


Консенсус FoxA :
5'-RYMAAYA-3'

[Overdier et al., 1994; Roux et al., 1995]

FoxA2 ChIP-seq в печени мыши [Wederell et al., 2008]

Число пиков– **11 475**
Средняя высота пика– **16**
Средняя длина пика– **727 bp**



•Wederell E.D. et al. Nucleic Acids Res. 2008 36: 4549–4564

PWM и SiteGA

Для создания обучающей выборки из 81 известных ССТФ FoxA использовали данные из базы TRRD и литературных источников.

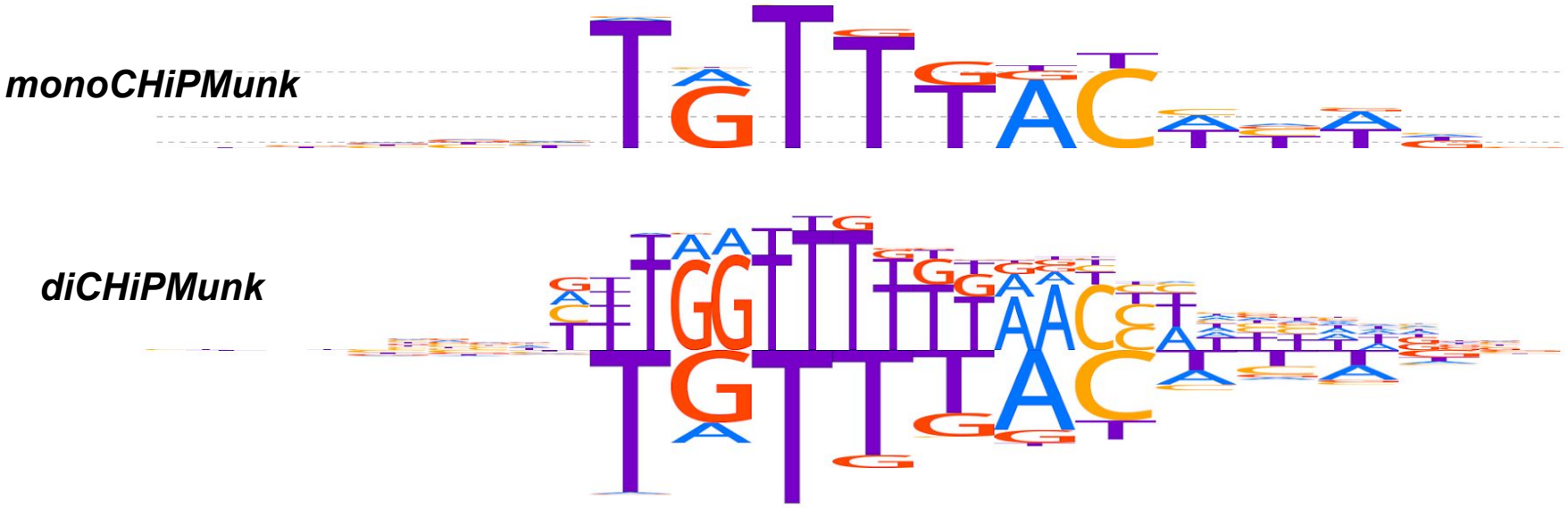
Критерием для отбора сайтов служило наличие доказательств взаимодействия FoxA с соответствующим районом ДНК, полученных одним из следующих методов:

- футпринтинг ДНКазой I с использованием очищенного белка;
- EMSA с использованием очищенного белка;
- EMSA с использованием ядерного экстракта белков и специфических антител (EMSA supershift).

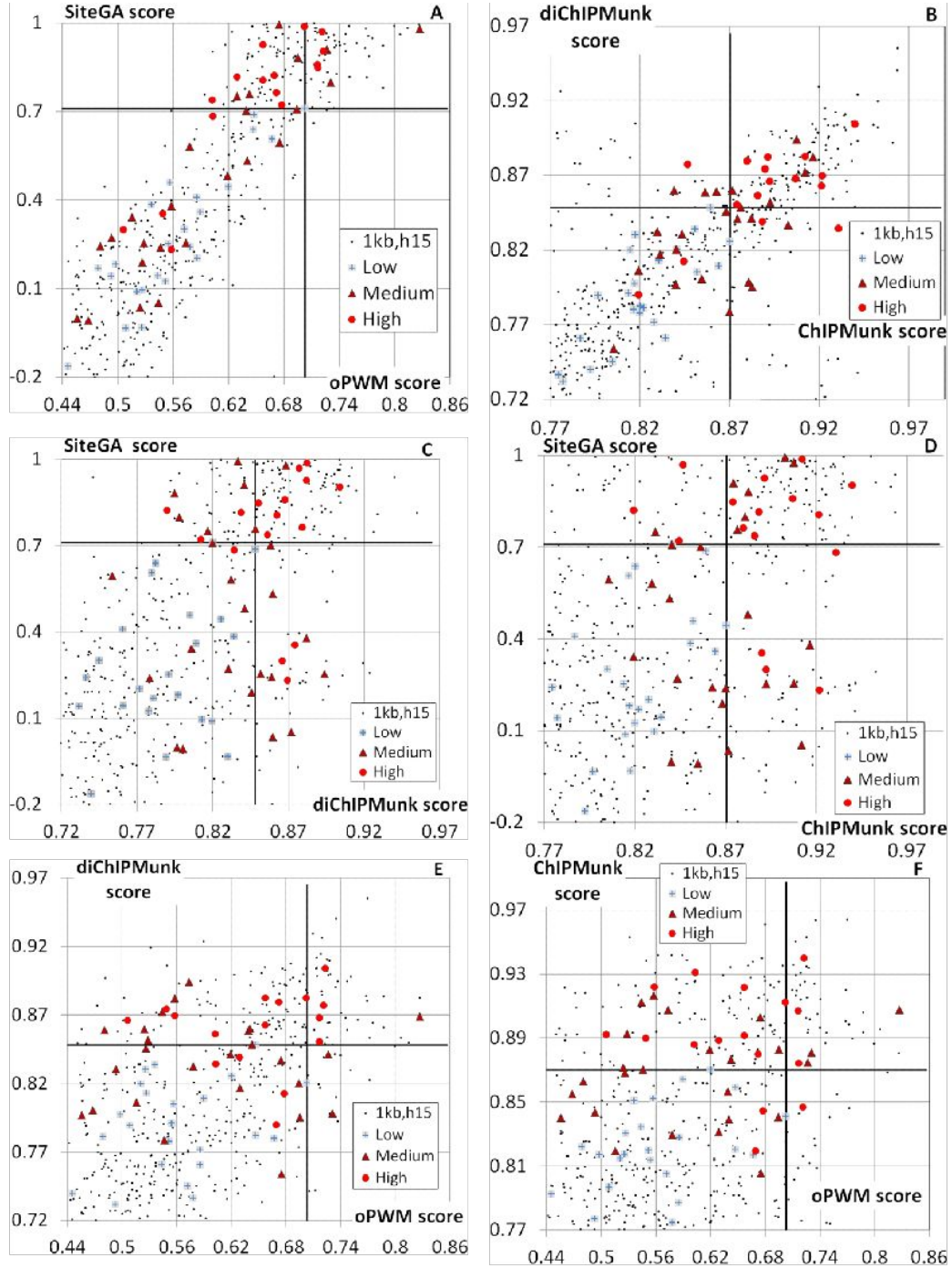
Выравнивание последовательностей проводилось относительно наиболее представленного мотива (в кодировке IUPAC) [Levitsky et al., 2011]. В итоге, для построения методов PWM и SiteGA использовалось выравнивание 53 последовательностей, содержащих мотив TRTTTRYH (R=A/G, Y=C/T, H=A/C/G).

CHiPMunk

Для обучения CHiPMunk были использованы 4455 локуса связывания FoxA2 (CHiP-seq; высота пика >15). Длины мотива для mono- & di-версий Chipmunk установлены как 20 нт и 28 нт, соответственно. Выявленные мотивы имеют высокий уровень сходства как с мотивом TRTTTRYH, полученным на выборке 81 CC FoxA, так и с известным консенсусом сайтов FoxA2 [Overdier et al., 1994; Roux et al., 1995].



Лого наиболее представленных мотивов, полученных с помощью mono- and diChipmunk для CHiP-seq FoxA2

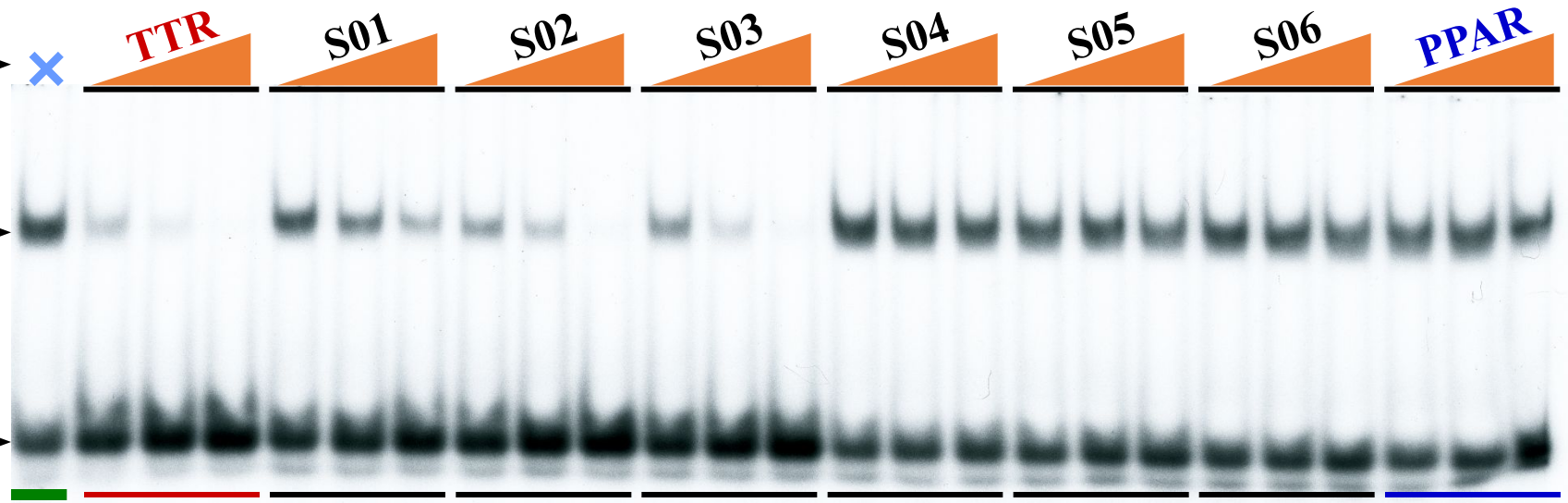

SiteGA score
A
1
0.7
0.4
0.1
-0.2
0.44 0.5 0.56 0.62 0.68 0.74 0.8 0.86
oPWM score
1kb,h15
Low
Medium
High
diChIPMunk score
B
0.97
0.92
0.87
0.82
0.77
0.72
0.77 0.82 0.87 0.92 0.97
ChIPMunk score
SiteGA score
C
diChIPMunk score
SiteGA score
D
ChIPMunk score
diChIPMunk score
E
0.97
0.92
0.87
0.82
0.77
0.72
ChIPMunk score
F
0.97
0.93
0.89
0.85
0.81
0.77
oPWM score

Экспериментальная верификация 64 предсказанных ССТФ FoxA методом задержки в геле (EMSA), конкурентный анализ

[^{32}P]-TTR: *меченый олигонуклеотид, соответствующий сайту связывания FoxA из промотора гена ttr мыши*

GST-DBD-FoxA2: *Задержка в геле GST-слитым белком, соответствующим ДНК-связывающему домену ТФ FoxA2*

немеченый конкурент: *×2, ×5 или ×20 молярный избыток конкурентного олигонуклеотида*



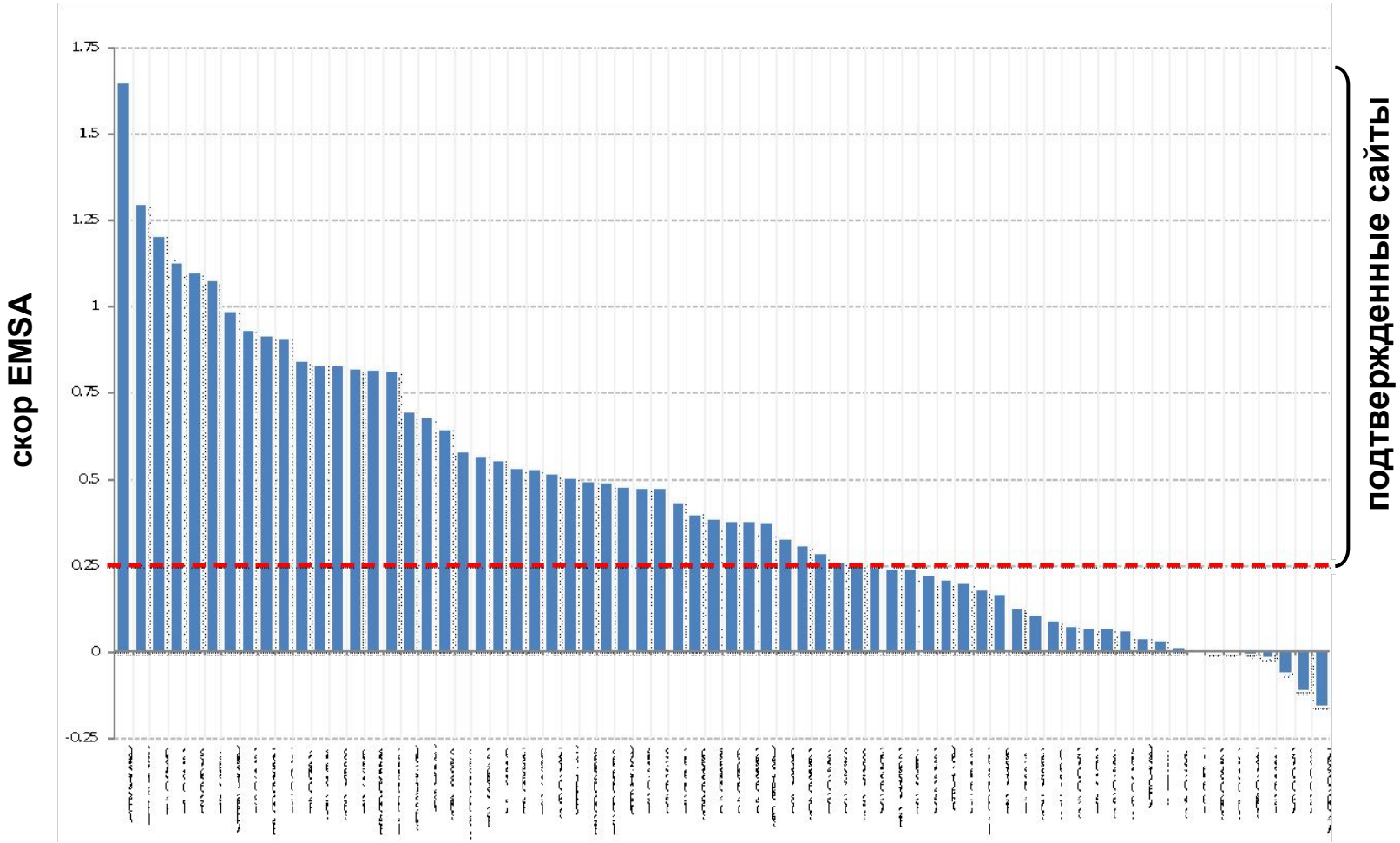
референс	положит. контроль	(0.78 | 0.15)	(0.81 | 0.70)	(0.79 | 0.92)	(0.74 | 0.44)	(0.77 | 0.26)	(0.79 | 0.13)	отрицат. контроль
		0.31	0.83	0.83	0.0	0.0	0.0	

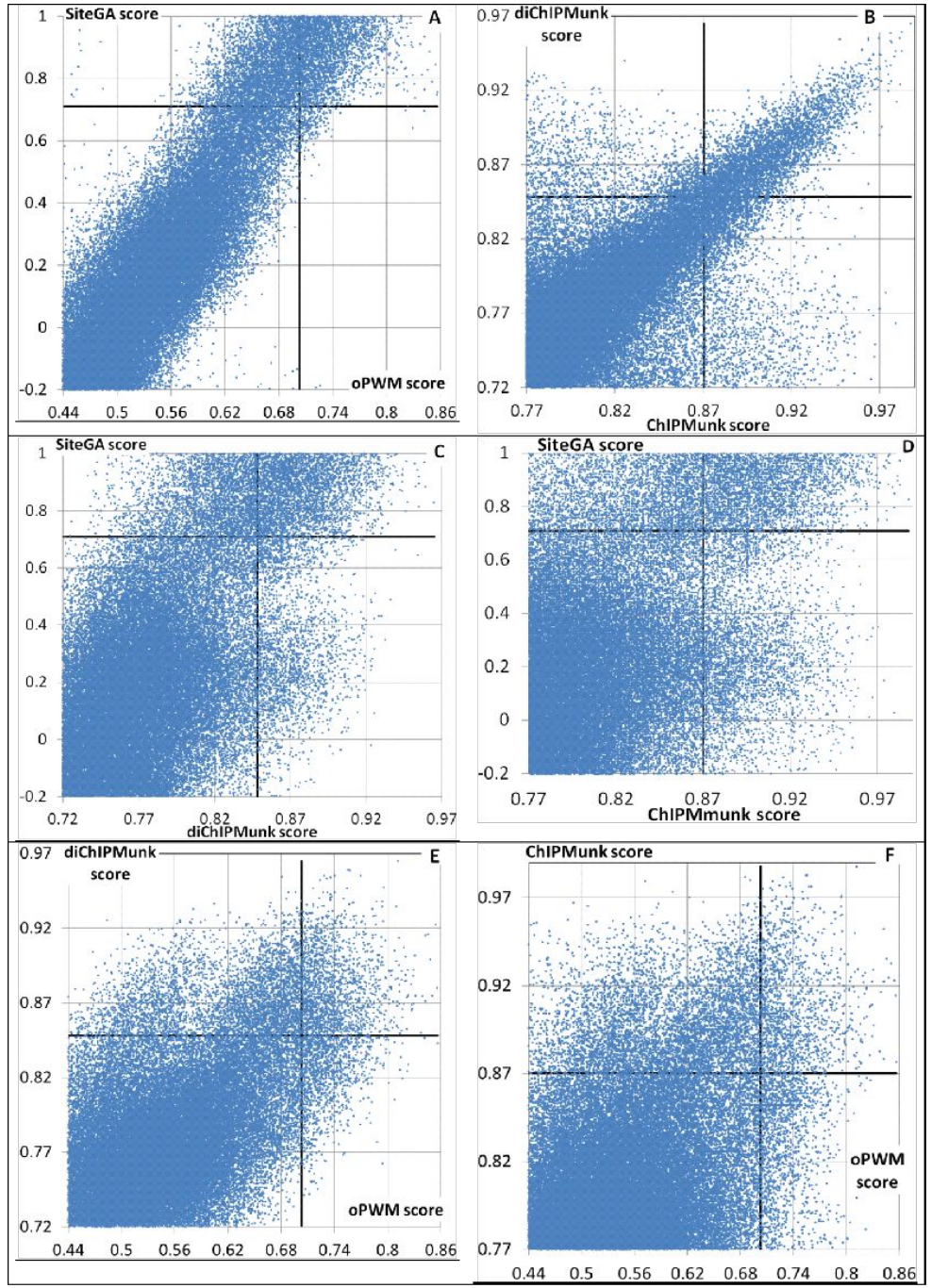
предсказанные ССТФ FoxA2

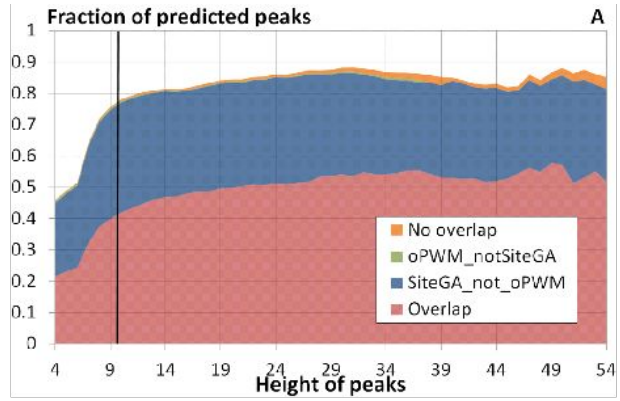
(скор CHiPMunk | скор SiteGA)
скор EMSA

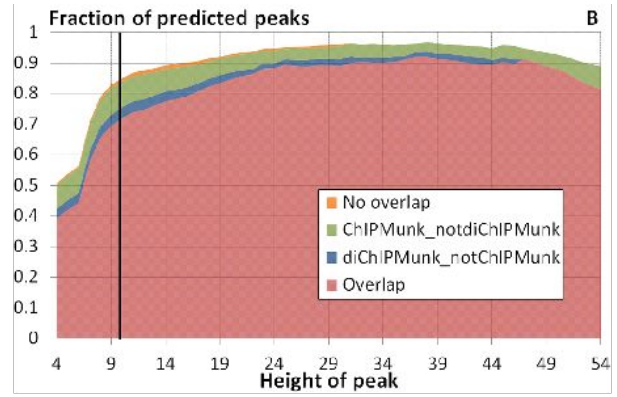
Экспериментальная верификация 64 предсказанных ССТФ FoxA методом задержки в геле (EMSA), конкурентный анализ

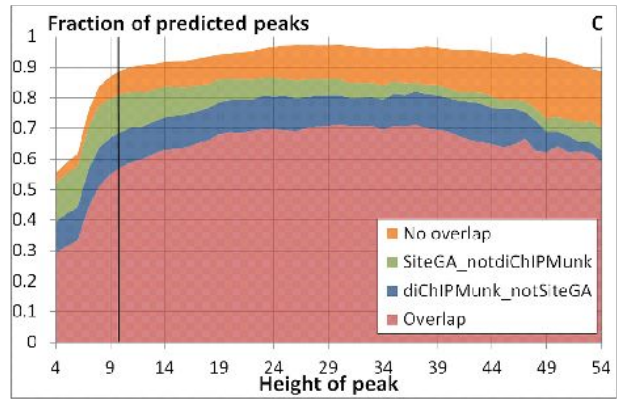
Скор EMSA расчитывался как отношение угла наклона линейной регрессии log-кривой зависимости от концентрации тестируемого олигонуклеотида к положительному контролю (TTR) и использовался в качестве оценки относительной аффинности этого олигонуклеотида. Поскольку скор EMSA распределен достаточно равномерно, порог отсечения неподтвержденных сайтов был выбран вручную (0.25).

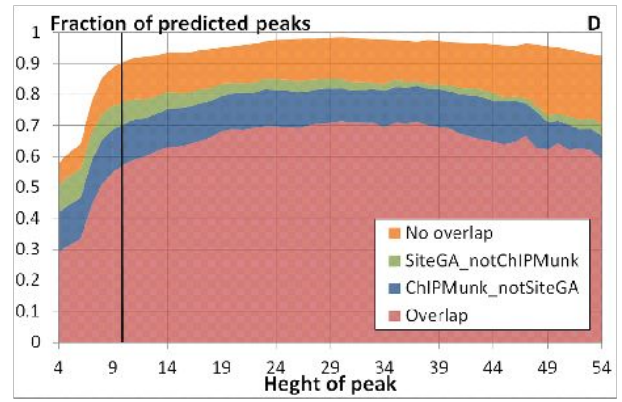


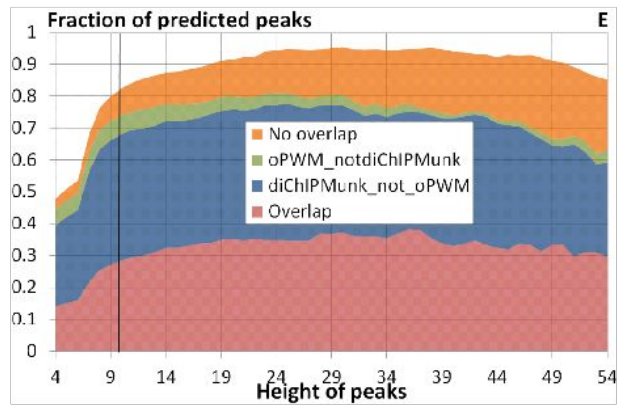

A
SiteGA score
oPWM score
1
0.8
0.6
0.4
0.2
0
-0.2
0.44 0.5 0.56 0.62 0.68 0.74 0.8 0.86
B
diChIPMunk score
ChIPMunk score
0.97
0.92
0.87
0.82
0.77
0.72
0.77 0.82 0.87 0.92 0.97
C
SiteGA score
diChIPMunk score
1
0.8
0.6
0.4
0.2
0
-0.2
0.72 0.77 0.82 0.87 0.92 0.97
D
SiteGA score
ChIPMmunk score
1
0.8
0.6
0.4
0.2
0
-0.2
0.77 0.82 0.87 0.92 0.97
E
diChIPMunk score
oPWM score
0.97
0.92
0.87
0.82
0.77
0.72
0.44 0.5 0.56 0.62 0.68 0.74 0.8 0.86
F
ChIPMunk score
oPWM score
0.97
0.92
0.87
0.82
0.77
0.44 0.5 0.56 0.62 0.68 0.74 0.8 0.86

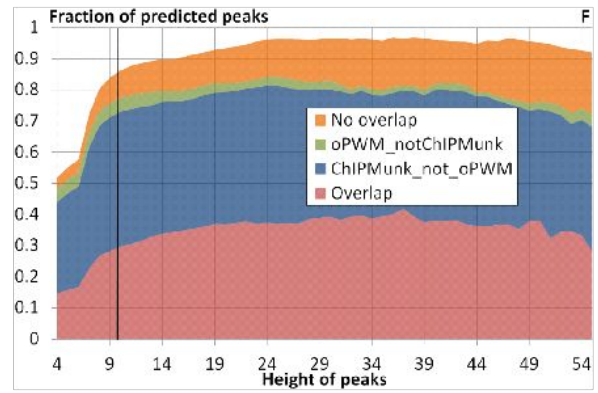

Fraction of predicted peaks
A
No overlap
oPWM_notSiteGA
SiteGA_not_oPWM
Overlap
Height of peaks


Fraction of predicted peaks
B
No overlap
ChIPMunk_notdiChIPMunk
diChIPMunk_notChIPMunk
Overlap
Height of peak


Fraction of predicted peaks
C
No overlap
SiteGA_notdiChIPMunk
diChIPMunk_notSiteGA
Overlap
Height of peak


Fraction of predicted peaks
D
No overlap
SiteGA_notChIPMunk
ChIPMunk_notSiteGA
Overlap
Heght of peak


Fraction of predicted peaks
E
No overlap
oPWM_notdiChIPMunk
diChIPMunk_not_oPWM
Overlap
Height of peaks


Fraction of predicted peaks
F
No overlap
oPWM_notChIPMunk
ChIPMunk_not_oPWM
Overlap
Height of peaks

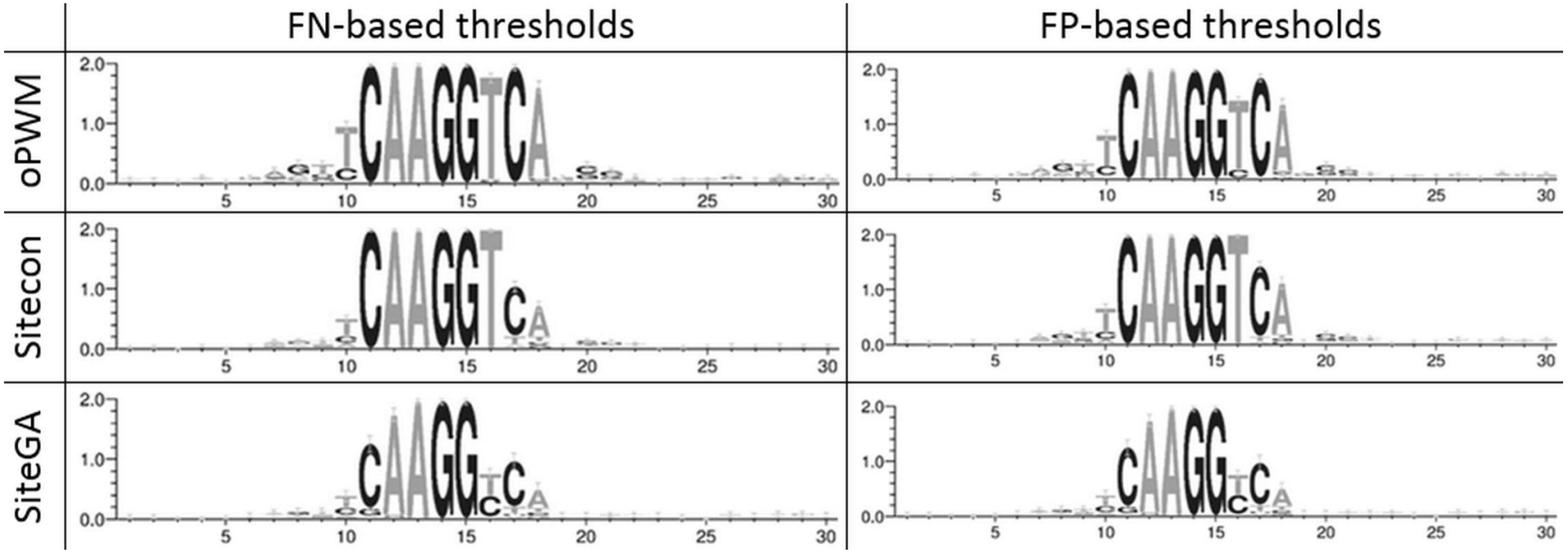
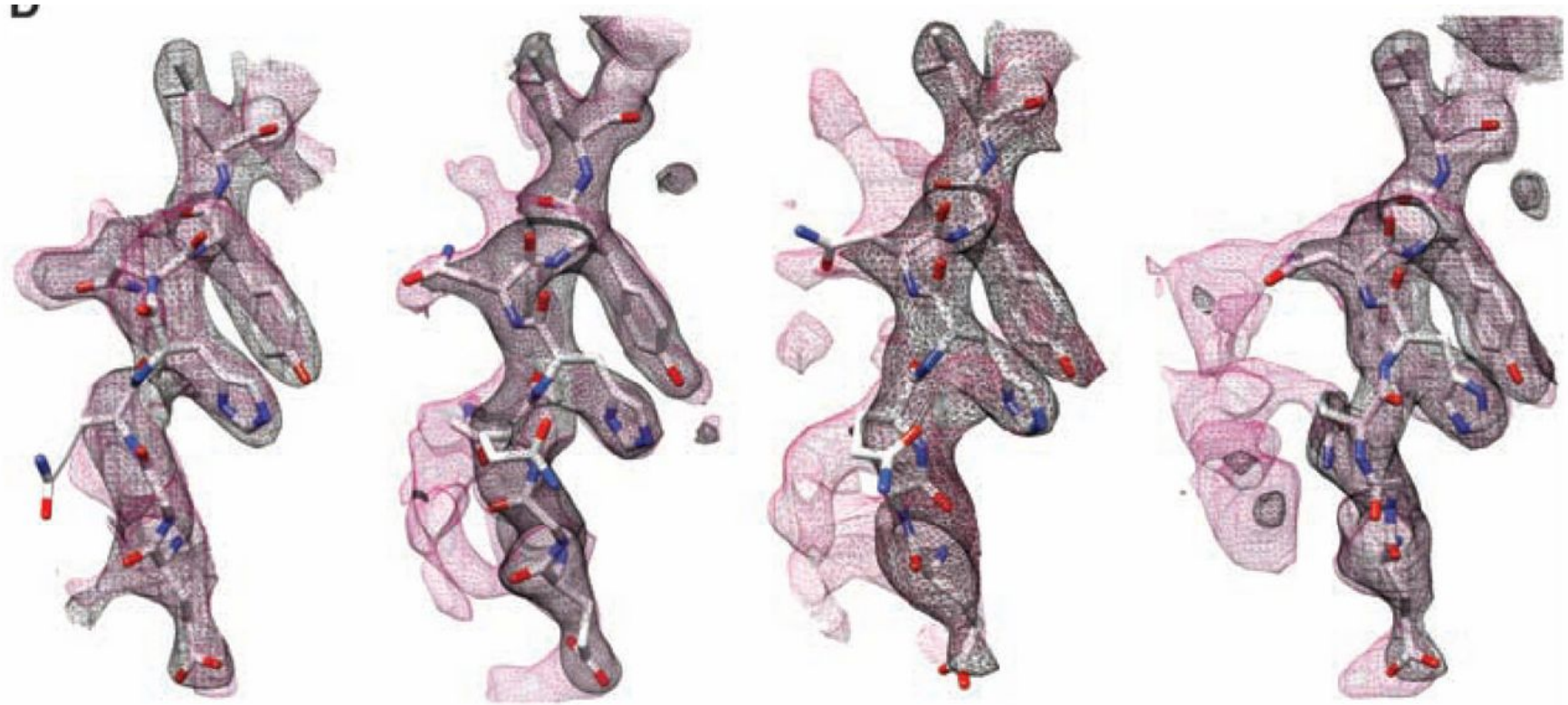
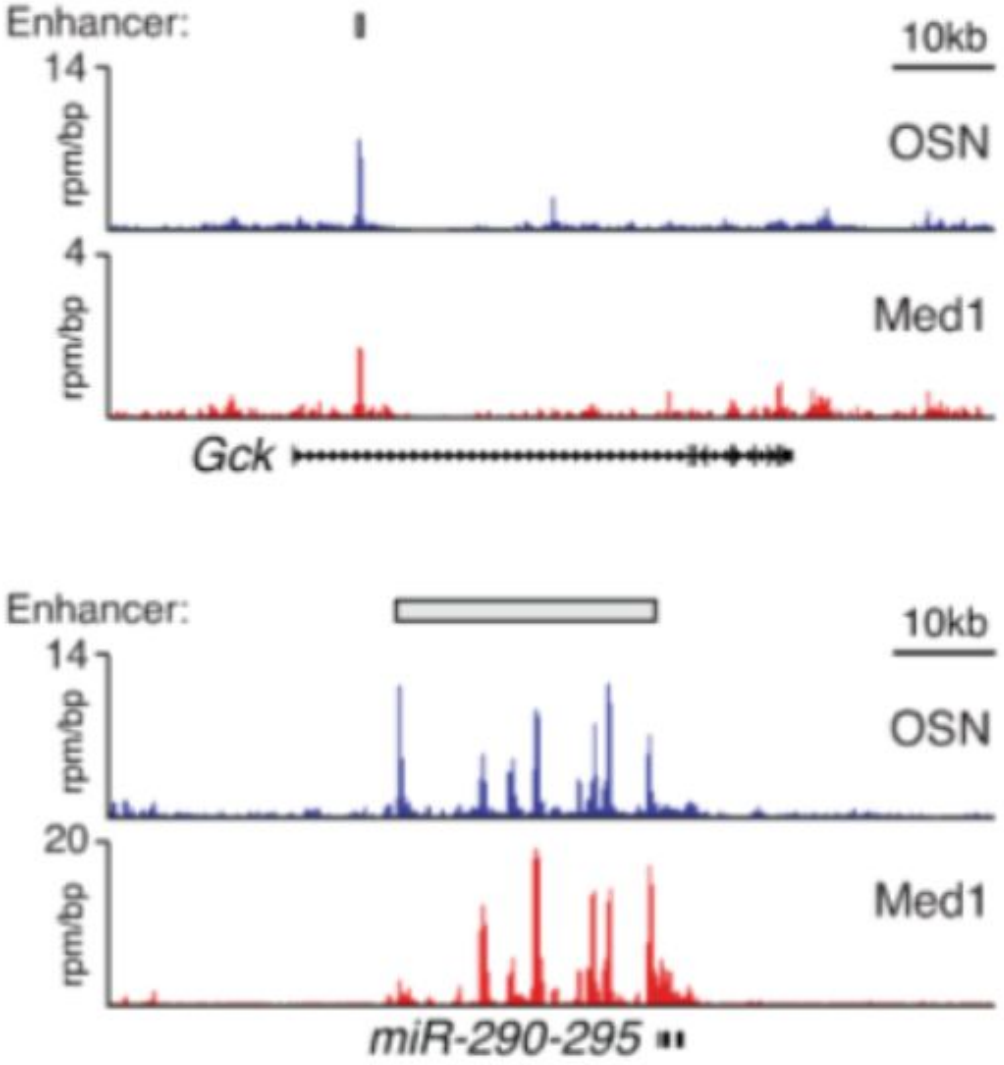


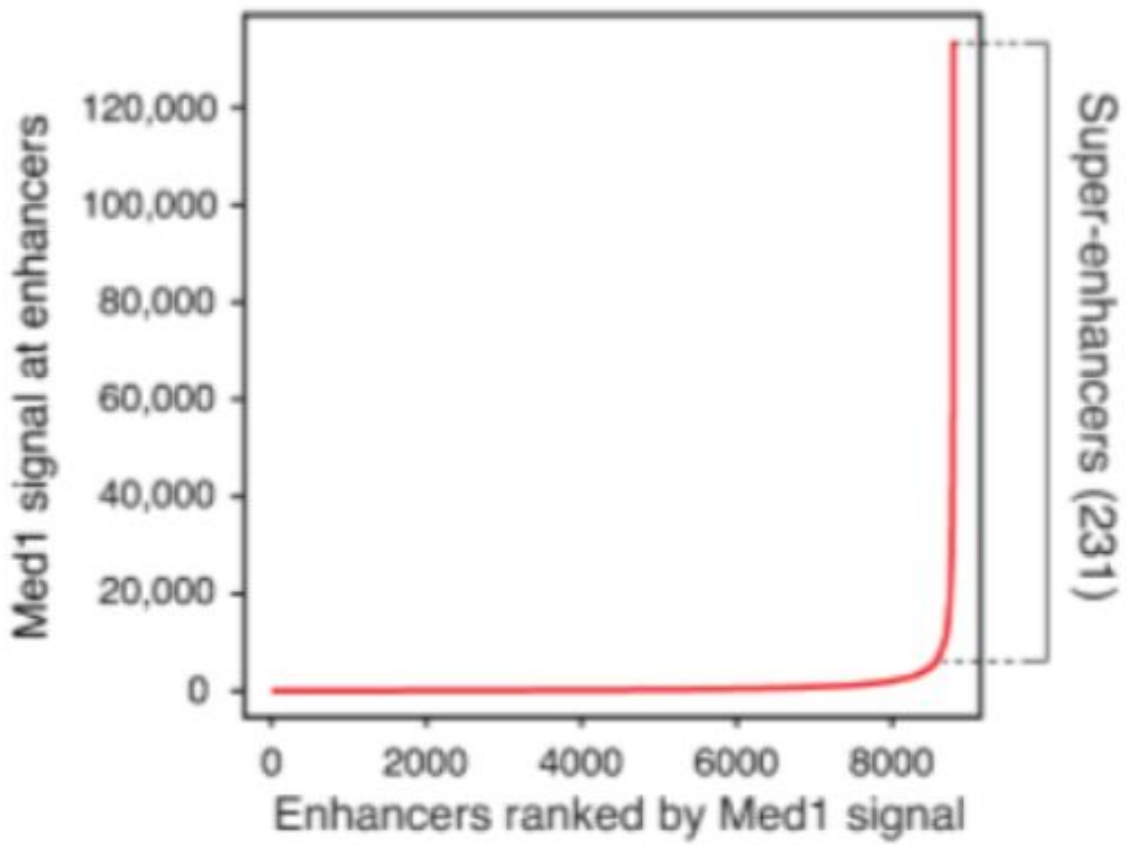
Figure 7. Context patterns of the SF-1 BSs recognized by sequence analysis of the top-scoring ChIP-seq peaks (peak height of 15 or higher). The frequency matrices were visualized by the WebLogo tool (Crooks et al., 2004) for the samples of sites predicted by oPWM, Sitecon, and SiteGA. The *X* axis shows nucleotide positions and *Y* axis, bits.

GBS	Sequence	Fold activation dex/etoh	K_D (μM)
Random/-	-	0.8±0.1	>10
GilZ	AGAACAttgGGTTCC	2.3±0.4	0.89±0.20
Pal	AGAACAaaaTGTTCT	2.0±0.1	0.08±0.01
Sgk	AGAACAtttTGTCCG	5.4±0.8	0.15±0.03
Tat	AGAACAtcccTGTACA	22.3±4.6	1.39±0.35
Cgt	AGAACAtttTGTACG	8.2±1.0	0.24±0.09
Cons	AGAACAaaaTGTACC	5.9±0.7	0.19±0.06
FKBP5	AGAACAgggTGTTCT	5.9±0.4	0.44±0.12

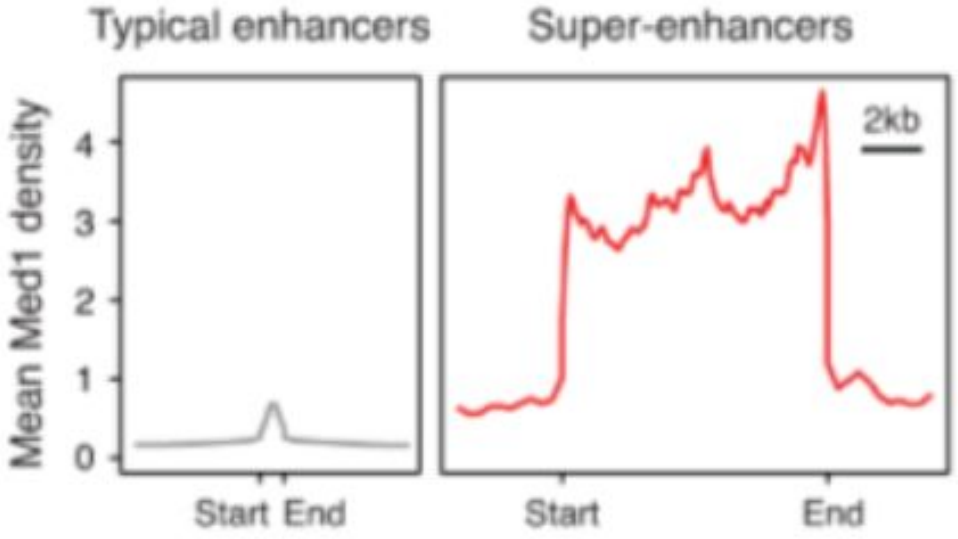

GR-DBD:GilZ
GR-DBD:Sgk
GR-DBD:FKBP5
GR-DBD:Pal

СУПЕРЭНХАНСЕРЫ

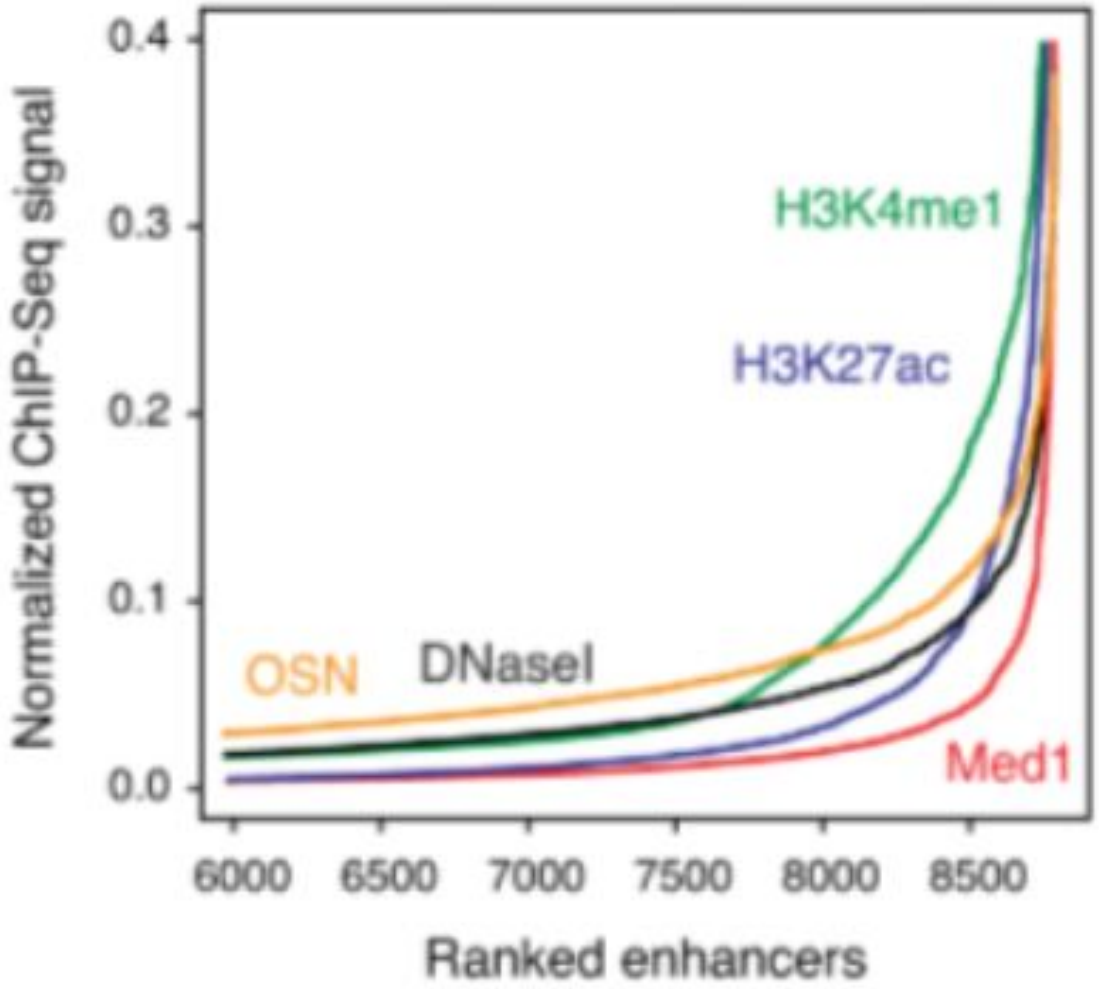




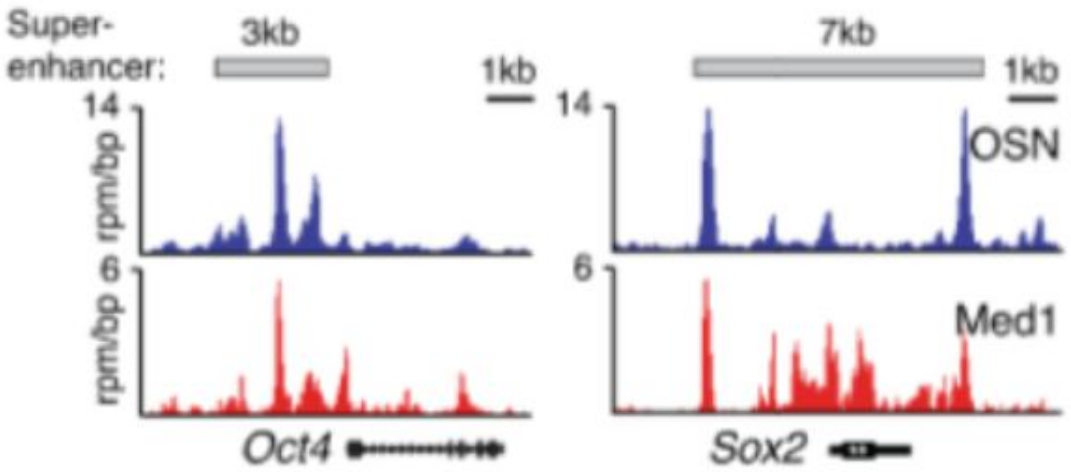
СУПЕРЭНХАНСЕРЫ



	Typical enhancers		Super-enhancers	
Number:	8563		231	
Median size:	703 bp		8667 bp	
	Total signal	Density at constituents	Total signal	Density at constituents
Med1:	1x	1x	28x	8.1x
H3K27ac:	1x	1x	26x	4.8x
H3K4me1:	1x	1x	10x	1.3x
DNaseI:	1x	1x	8x	2.2x



СУПЕРЭНХАНСЕРЫ



Selected genes associated with super-enhancers

Transcription factors

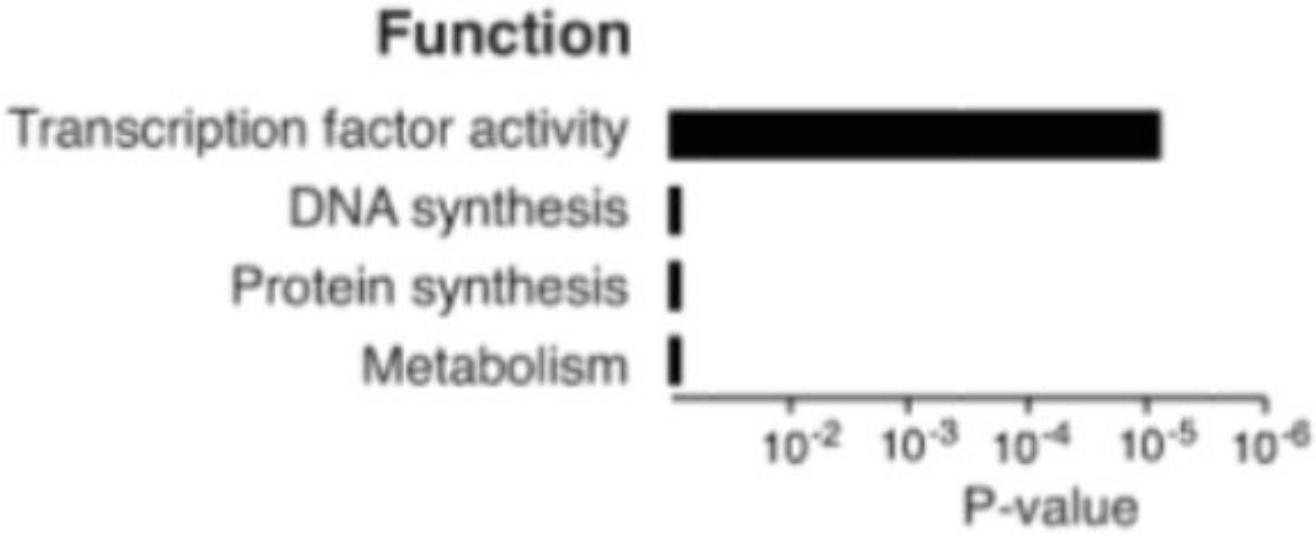
Oct4, Sox2, Nanog, Klf4, Esrrb, n-Myc, Utf1, Sall4, Prdm14, Dppa5a, Tbx3, Zfp42

Chromatin modifying enzymes

Tet1, Tet2

Micro RNAs

miR-290-295



СУПЕРЭНХАНСЕРЫ