Биоинформатика ДНК-белковых взаимодействий

До 90 % белков млекопитающих сходны - почему организмы различны?

- Чем отличаются последовательности ДНК разных видов?
- Почему сходные гены работают по разному?

Основная проблема биологии

- Почему конкретные гены работают в конкретный момент времени (в конкретной ткани)

За счет регуляции синтеза РНК и последующих стадий

- Процессы на уровне РНК (сплайсинг и др.)
- Экспорт РНК из ядра
- Трансляция
- Модификация белков

Изучаются разными методами

- Процессы на уровне РНК (сплайсинг и др.) - RNA - Seq
- Экспорт РНК из ядра -
- Трансляция - Ribo-Seq
- Модификация белков - MS

Лазерная микродиссекция позволяет брать образцы небольшого количества клеток

Primary cells

Тканеспецифичная экспрессия

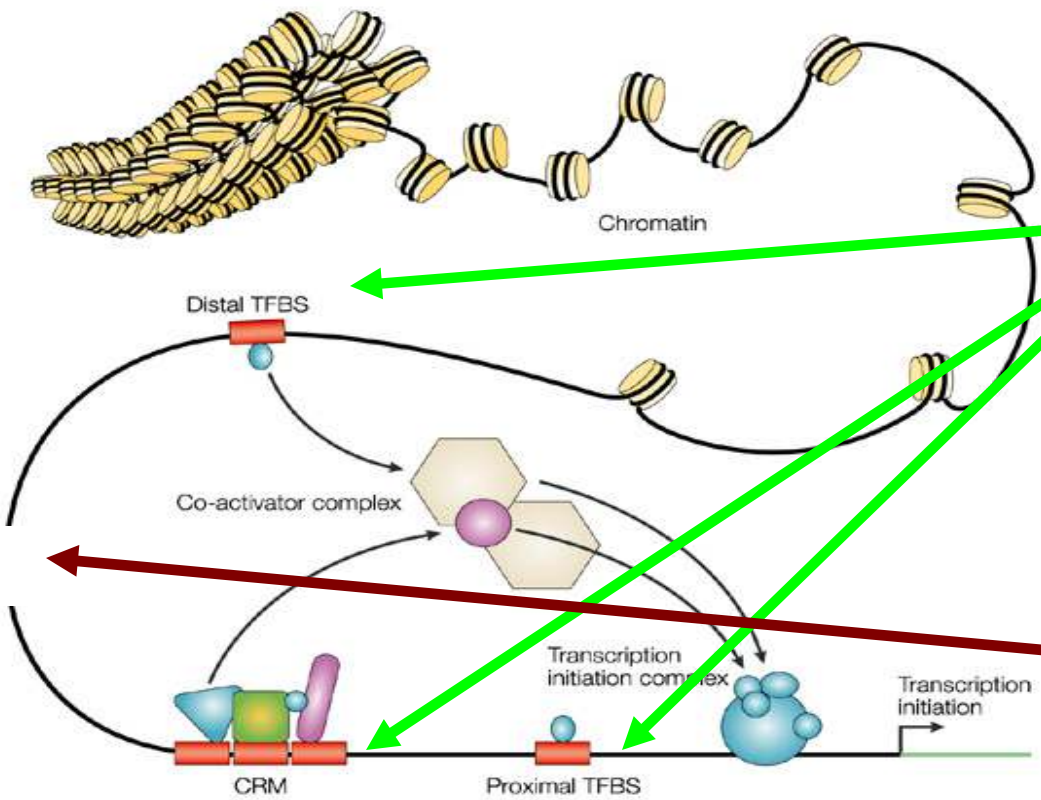
- Истинно генов домашнего хозяйства - мало (~8%)
- В зрелой ткани редко экспрессируется больше, чем 10-15% генов
- В эмбриональных тканях - больше
- В раковых тканях - совсем много

Регуляция транскрипции

- Изучена лучше всего (методически проще - можно манипулировать с ДНК и РНК)
- Успехи и разочарования

ДНК текст: проекция 3D на 1D

Функциональная информация (промоторы и т.д.) неоднозначна!



Wasserman, W.W. and Sandelin, A. (2004) Nature Reviews | Genetics

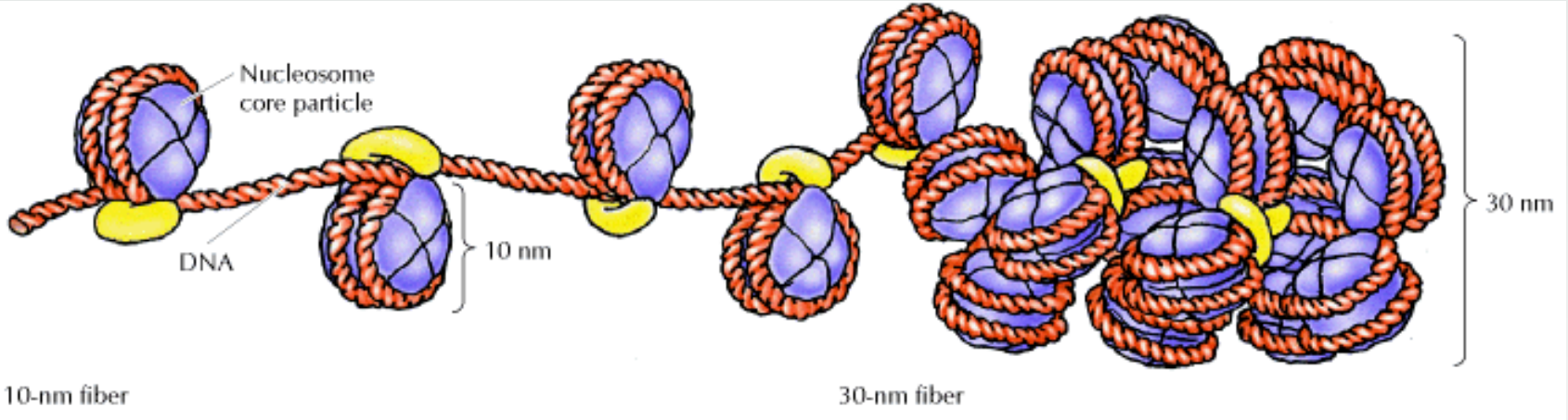
В эукариотах несколько сегментов ДНК контролируют экспрессию одного гена

Регуляторные области могут быть удалены до 100 кб от гена и даже могут быть на другой хромосоме

Регуляторы транскрипции…

- Обычно садятся на «открытый хроматин» (он же доступен к расщеплению ДНКазой)
- Хромосомы влияют по разному - связыванию некоторых факторов они препятствуют, а связыванию других - способствуют

Регуляция экспрессии генов осуществляется на уровне ДНК….

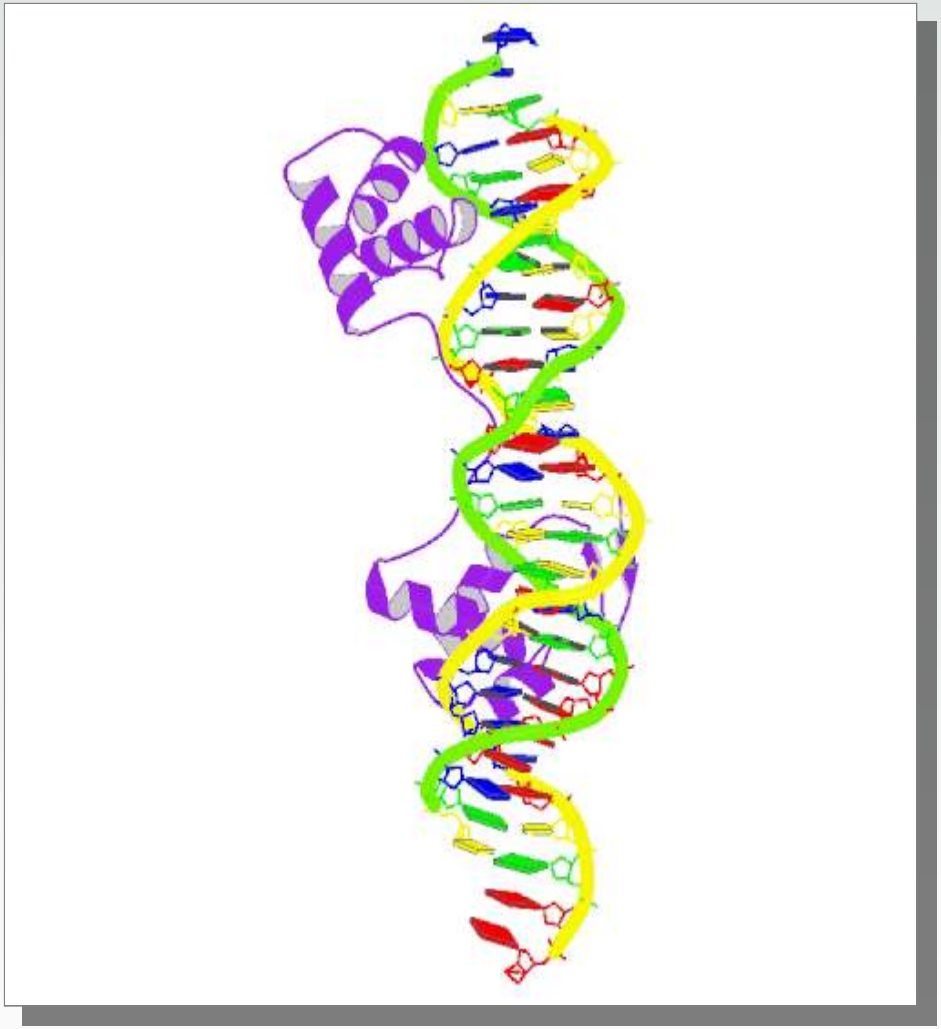


Белки взаимодействуют с ДНК

Около 1500 факторов у человека

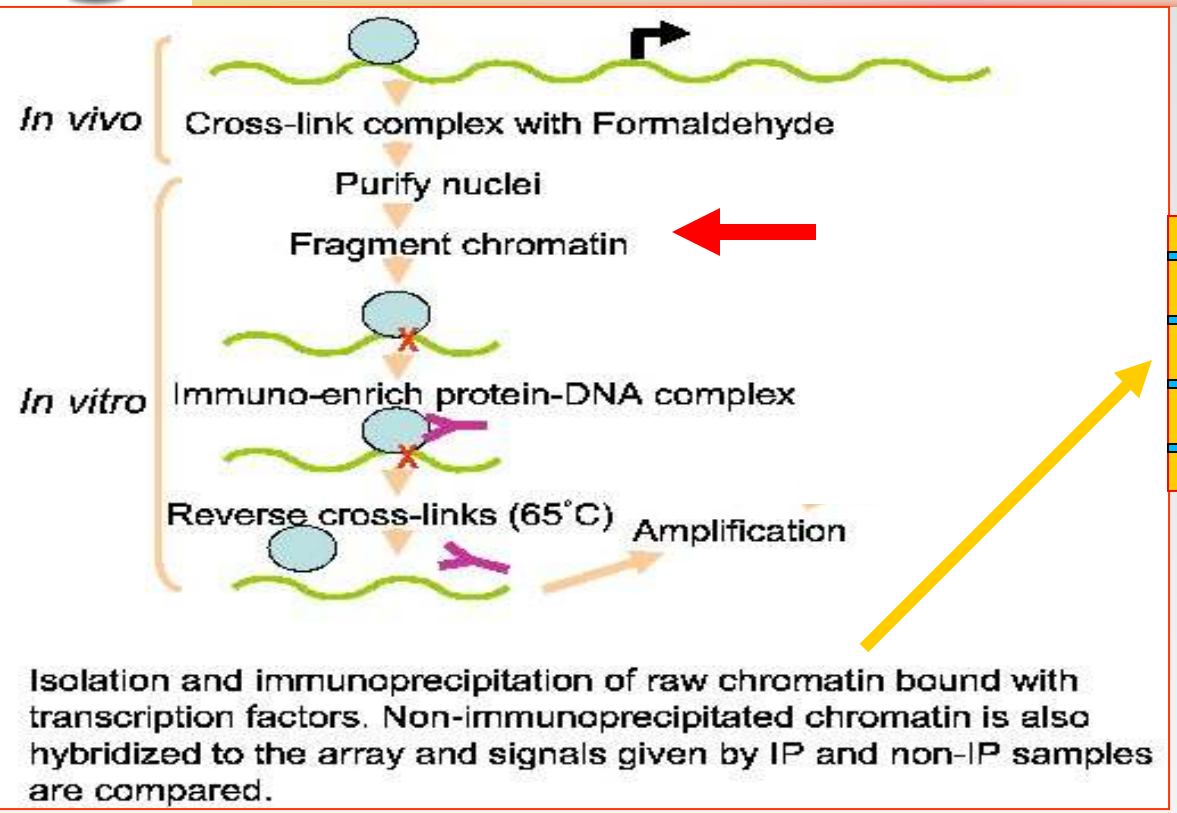
Изучено как-то – 50%

Изучено хорошо – 200шт



From E. Wingender, BGRS 2012

Иммунопреципитация хроматина (ChIP)

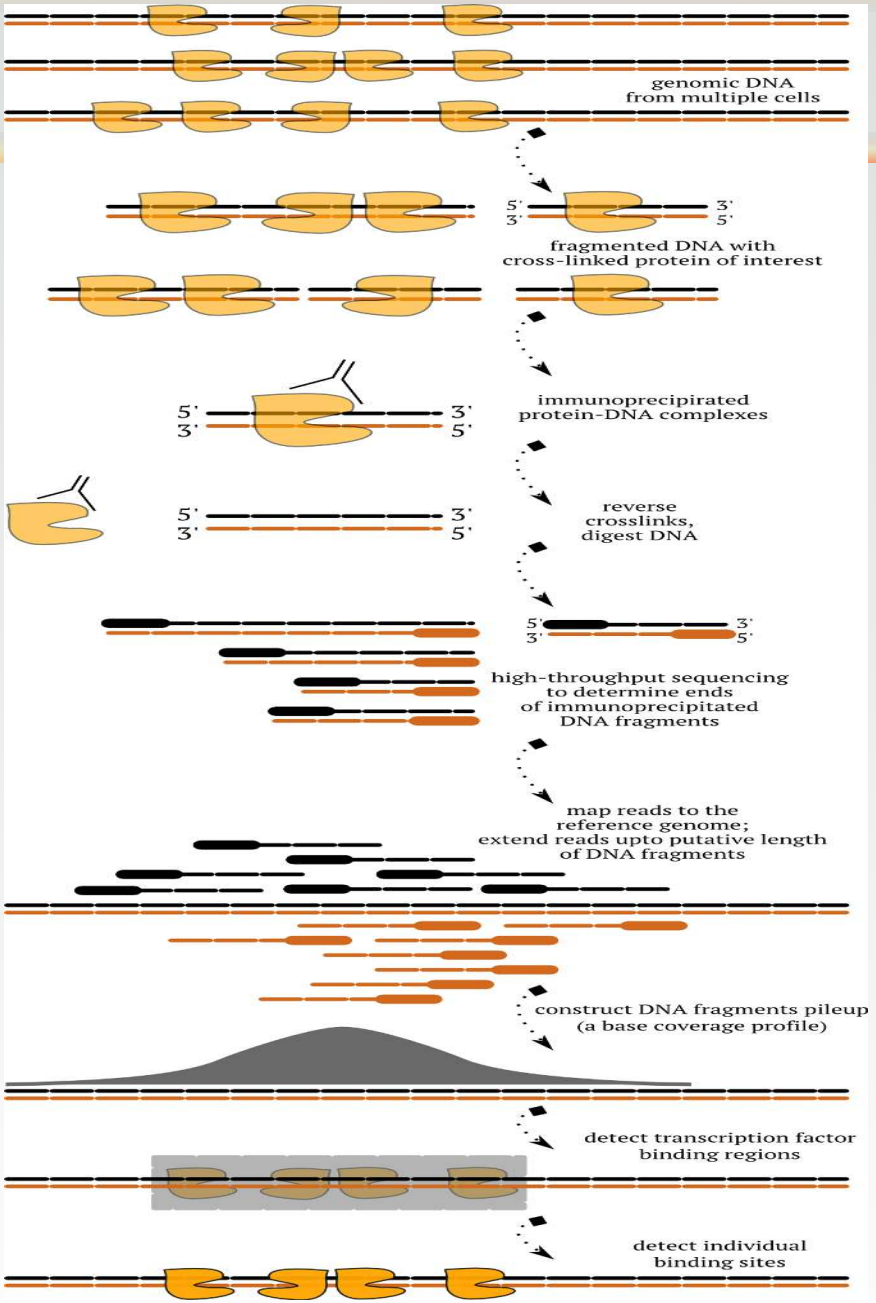


- Гибридизация ChIP-chip
- Секвенирование ChIP-seq
- Замыкание ChIP-PET
- Обработка экзонуклеазой ChIP-exo

Из: http://www.tigr.org/

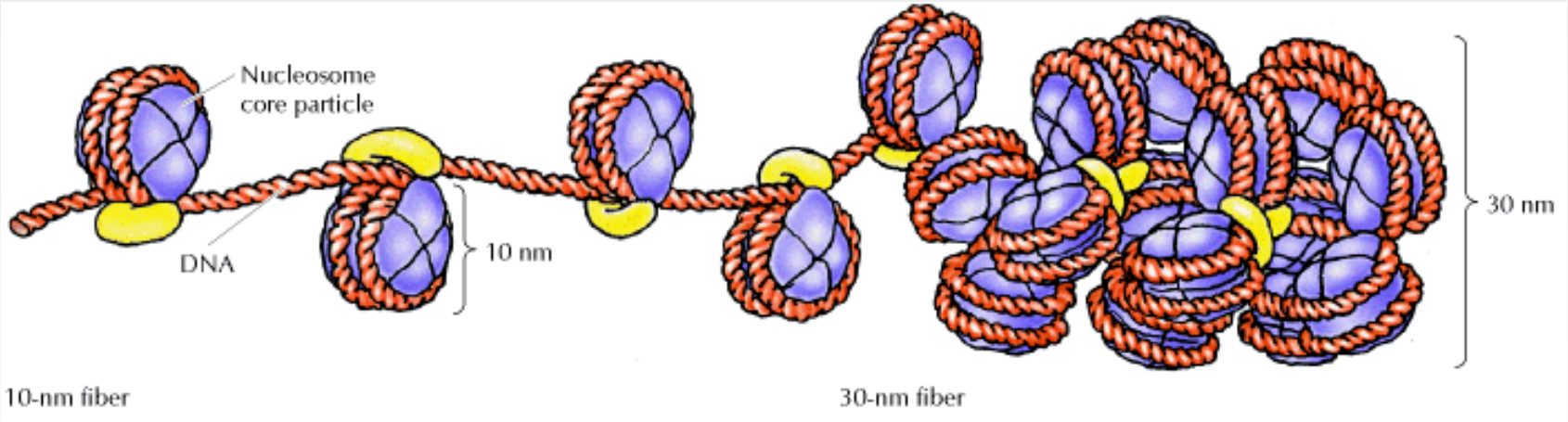
Достоинство: Полногеномный анализ связывания *in vivo*

Недостаток: Фактически изучается связывание в фрагментах не короче ~300по, за исключением ChIP-exo

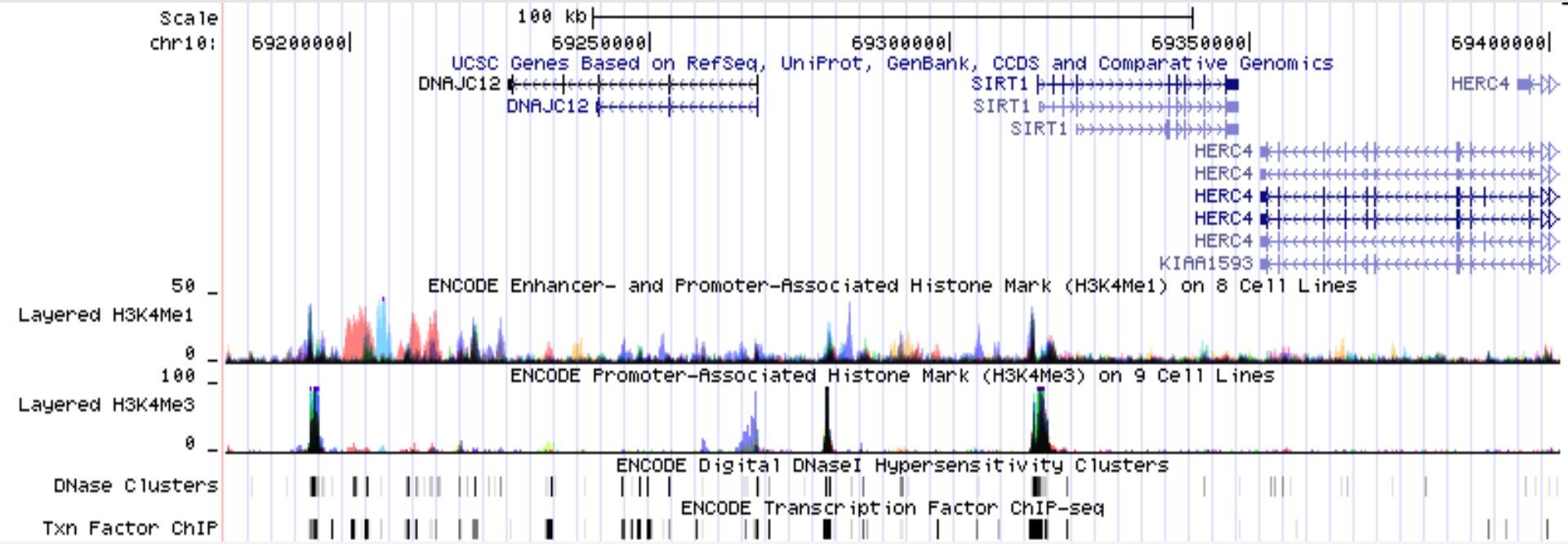

genomic DNA
from multiple cells
5'
3'
3'
5'
fragmented DNA with
cross-linked protein of interest
immunoprecipitated
protein-DNA complexes
reverse
crosslinks,
digest DNA
high-throughput sequencing
to determine ends
of immunoprecipitated
DNA fragments
map reads to the
reference genome;
extend reads upto putative length
of DNA fragments
construct DNA fragments pileup
(a base coverage profile)
detect transcription factor
binding regions
detect individual
binding sites

ChIP - видит не только ТФ

- можно смотреть нуклеосомы, в частности их химические модификации
- в зависимости от фукнций участка ДНК - нуклеосомы по разному химически меняются
- можно делать антитела на химически изменненные нуклеосомы и изучать их тем же образом

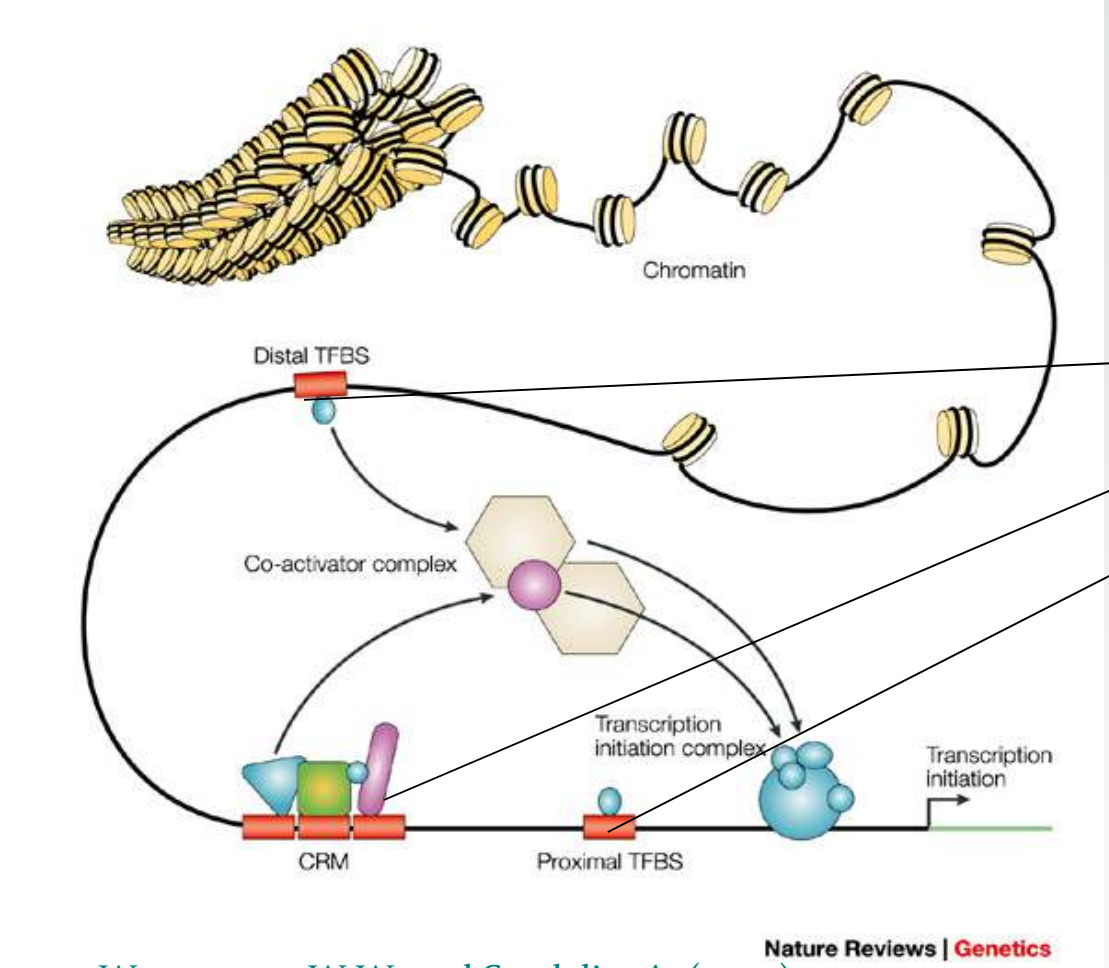


UCSC


Scale
chr10:
100 kb
69200000
69250000
69300000
69350000
69400000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
DNAJC12
DNAJC12
SIRT1
SIRT1
SIRT1
HERC4
HERC4
HERC4
HERC4
HERC4
HERC4
HERC4
KIAA1593
50
Layered H3K4Me1
0
ENCODE Enhancer- and Promoter-Associated Histone Mark (H3K4Me1) on 8 Cell Lines
100
Layered H3K4Me3
0
ENCODE Promoter-Associated Histone Mark (H3K4Me3) on 9 Cell Lines
DNase Clusters
ENCODE Digital DNaseI Hypersensitivity Clusters
Txn Factor ChIP
ENCODE Transcription Factor ChIP-seq

ССТФ и анализ последовательностей ДНК

Участки ДНК, связывающие регуляторные белки содержат похожие мотивы ДНК



Wasserman, W.W. and Sandelin, A. (2004)

CCCGACCA

AAATTTAT
(piggy back)

CCCGACTA

Некоторые из участков ДНК, согласно экспериментальными данным связывающие конкретный ТФ, содержат похожие слова. Такие слова встречаются и в участках ДНК, не связывающих этот ТФ.

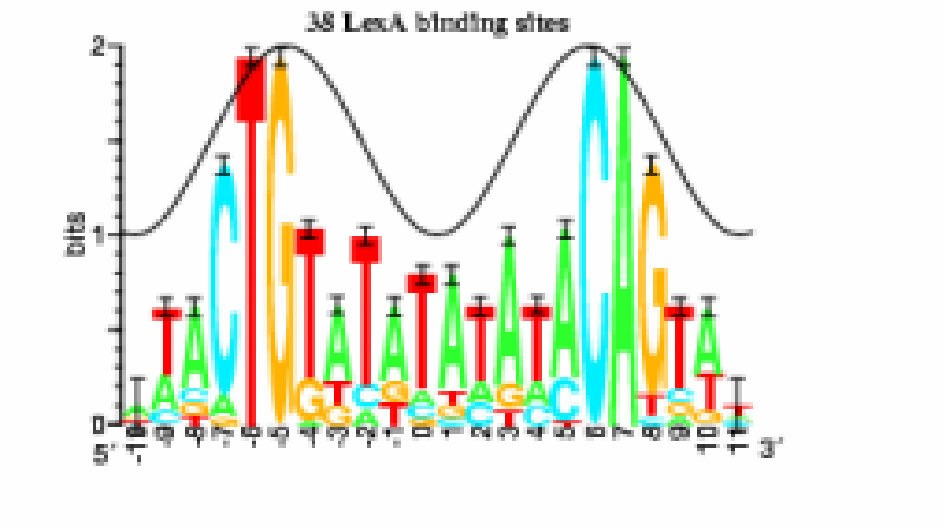
Результаты экспериментов по связыванию белков на ДНК - сегменты разной длины

Метод	In vivo/ in vitro	**Природная/ Синтези-рованная ДНК**	**Длина активного сегмента**	**Число активных сегментов**	**Недостатаки Комментарии**
ChIP {Seq, Pet, Chip}	In vivo	**Природная**	500 - 5000	500 – 50 000	Непрямое связывание (Piggy back)
Одногибридная система	in vivo **(модельные виды)**	**Синтетич.**	28	20-50	Связывание в бактериях
SELEX, RSS	In vitro	**Синтетич.**	20	20-50	Структура олига
PBA	In vitro	**Синтетич.**	60	>10000	Де Брюин последовательность
Футпринты	In vitro	**Природная**	20-100	5-30	Сдвиг защищенной области
HP-Footprint	In vivo	**Природная**	100-200	200000	Плохое разрешение факторов
HP-Selex	In vitro	**Синтетич.**	10-20	20-50	Структура олига, насыщение циклов

Методы in vitro позволяют контролировать связывание конкретного белка

Позиции контактов более консервативны, как в ДНК так и в белке

TACTGTATATATATACAGTA	Site1
ACCTGGTTACGTACACAGCT	Site2
TAATGTATATATATACATTA	Site3
AGCTGTACTTAAGTACAGCT	Site4
TTCTGGGAGCGCGACCAGGA	Site5

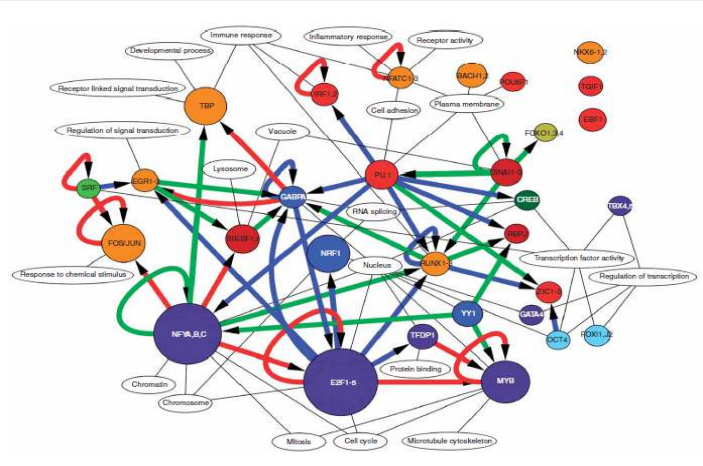


Что нам дает знание где расположены ССТФ?

...AGACTTCCACGAACTTTA...

Определение важных точек в ДНК (rSNP, remodeling)

Возможность предсказывать регуляторные сети



Suzuki, Forrest (2009)
Nature genetics

Модели участков связывания

A	G	A	T	A	A
T	G	A	T	A	G
T	G	A	T	A	C
G	G	A	T	A	C
G	G	A	T	T	T

Initial set of aligned sequences

	1	2	3	4	5	6
1	A	G	A	T	A	A
2	T	G	A	T	A	G
3	T	G	A	T	A	C
4	G	G	A	T	A	C
5	G	G	A	T	T	T

	1	2	3	4	5	6
A	1	0	5	0	4	1
C	0	0	0	0	0	2
G	2	5	0	0	0	1
T	2	0	0	5	1	1

Position Count/Frequency Matrix (PCM/PFM)

LOGO representation

	1	2	3	4	5	6
A	0.2	0.0	1.0	0.0	0.8	0.2
C	0.0	0.0	0.0	0.0	0.0	0.4
G	0.4	1.0	0.0	0.0	0.0	0.2
T	0.4	0.0	0.0	1.0	0.2	0.2

Position Probability Matrix (PPM)

	1	2	3	4	5	6
A	-0.2	-1.8	1.2	-1.8	1.0	-0.2
C	-1.8	-1.8	-1.8	-1.8	-1.8	0.4
G	0.4	1.2	-1.8	-1.8	-1.8	-0.2
T	0.4	-1.8	-1.8	1.2	-0.2	-0.2

Position Weight Matrix (PWM)

Position Specific Scoring Matrix (PSSM)

position

weight

nucleotide letter

The motif model

Positional Weight Matrix: well-known classic model

[von Hippel, 1987]

PWM GGATTA

	1	2	3	4	5	6
A	-1.6	-1.6	0.96	-1.6	-1.6	0.96
C	-1.6	-1.6	**0.00**	-1.6	-1.6	-1.6
G	1.22	1.22	-1.6	-1.6	-1.6	-1.6
T	-1.6	-1.6	-1.6	1.22	1.22	**0.00**

$S_{GGATTA}=1.22+1.22+0.96+1.22+1.22+0.96=6.8$

$N=4$ { GGATTA, GGATTA, GGATTT, GGCTTA } width=6

word list → Positional Count Matrix → additon of the pseudocount → Positional Weight Matrix

	j 1	2	3	4	5	6
a A	0	0	3	0	0	3
C	0	0	1	0	0	0
G	4	4	0	0	0	0
T	0	0	0	4	4	1

$x_{\alpha,j}$

	1	2	3	4	5	6
A	0.25	0.25	3.25	0.25	0.25	3.25
C	0.25	0.25	1.25	0.25	0.25	0.25
G	4.25	4.25	0.25	0.25	0.25	0.25
T	0.25	0.25	0.25	4.25	4.25	1.25

$x_{\alpha,j}+aq_\alpha$

$q_\alpha=0.25$ background frequency

$a=1$ pseudocount

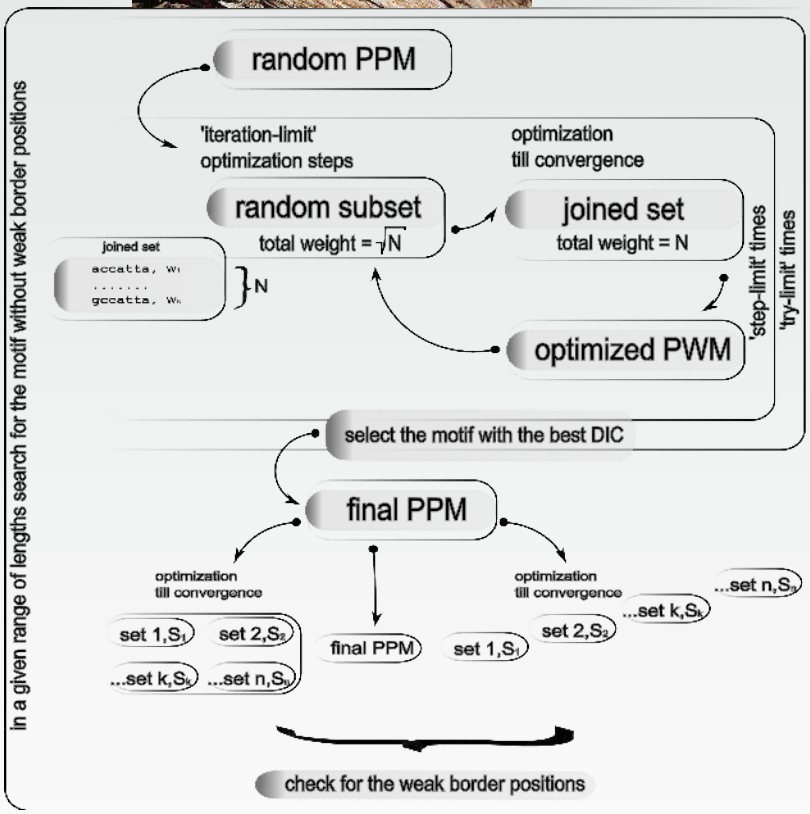
	1	2	3	4	5	6
A	-1.6	-1.6	0.96	-1.6	-1.6	0.96
C	-1.6	-1.6	**0.00**	-1.6	-1.6	-1.6
G	1.22	1.22	-1.6	-1.6	-1.6	-1.6
T	-1.6	-1.6	-1.6	1.22	1.22	**0.00**

$$S_{\alpha,j}=\ln\left(\frac{x_{\alpha,j}+aq_\alpha}{(N+a)q_\alpha}\right)$$

Выравнивание сайтов и построение модели

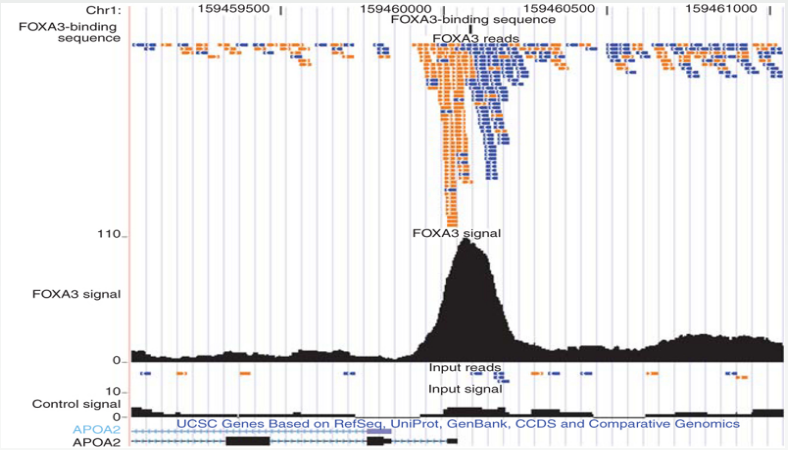
- Связанные процессы - должны быть итерации

ChIPmunk : Использование данных in vivo и in vitro для повышения надежности информации о связывании



http://seqanswers.com/wiki/ChIPMunk

- Эвристика, основанная на подвыборках исходных данных;
- позволяет интегрировать данные из разных экспериментальных источников с помощью весовых коэффициентов;
- использует априорную информацию о положении мотива.

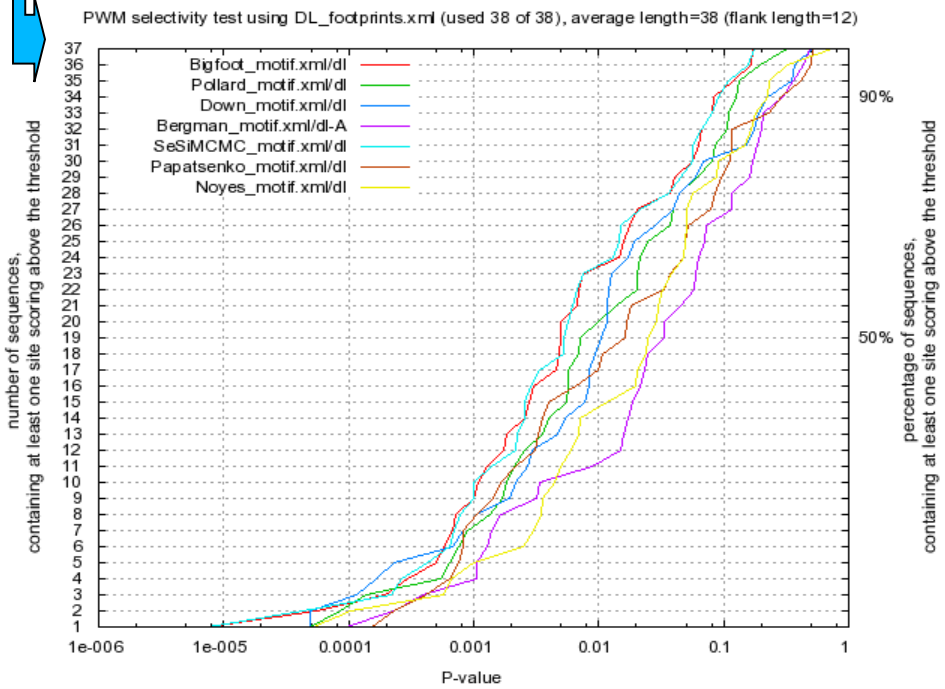


Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ
Bioinformatics (2010), 26(20), 2622

Разные технологии дают сходные, но отличающиеся характерные последовательности

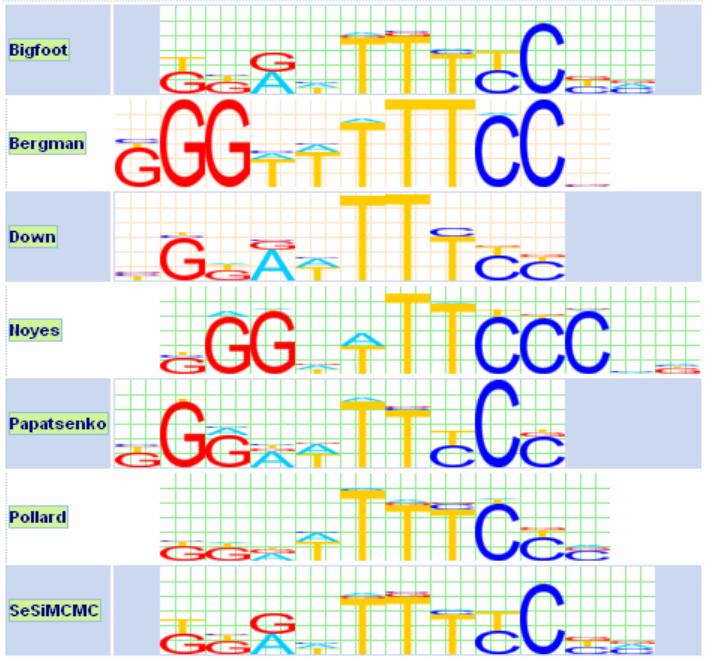
<up> DMMPMM motif DL selectivity testing [dmmpmm_compare_html]

футпринтов



% всех слов в модели

motif alignment



motif similarity

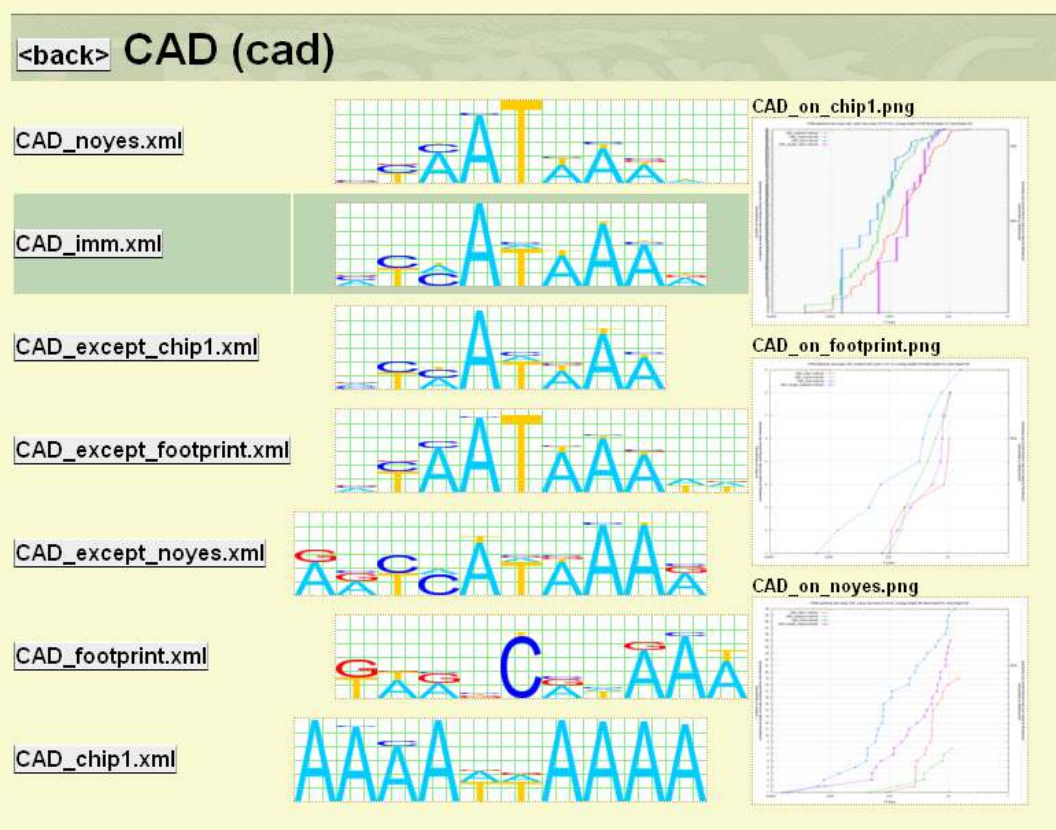
	Bigfoot	Pollard	Down	Bergman	SeSiMCMC	Papatsenko	Noyes
Bigfoot	1.0	0.3913	0.3124	0.2659	0.9566	0.3747	0.3158
Pollard	0.3913	1.0	0.4562	0.3381	0.3977	0.4415	0.4317
Down	0.3124	0.4562	1.0	0.2963	0.3106	0.3935	0.3208
Bergman	0.2659	0.3381	0.2963	1.0	0.2681	0.3575	0.3606
SeSiMCMC	0.9566	0.3977	0.3106	0.2681	1.0	0.3834	0.3206
Papatsenko	0.3747	0.4415	0.3935	0.3575	0.3834	1.0	0.4437
Noyes	0.3158	0.4317	0.3208	0.3606	0.3206	0.4437	1.0

Пример: дрозофила

Источники данных:

- SELEX [C.Bergman, D.Papatsenko]
- ChIP-chip [M.Biggin]
- Footprinting [C.Bergman]
- Bacterial one-hybrid system [M.Noyes]

Для 34 ТФ есть данные из более чем двух источников

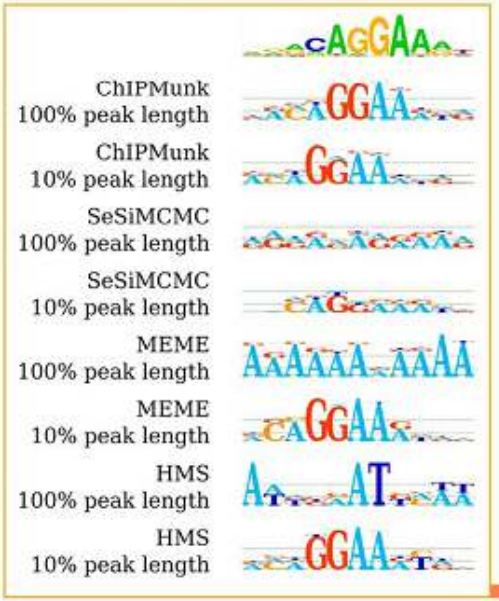


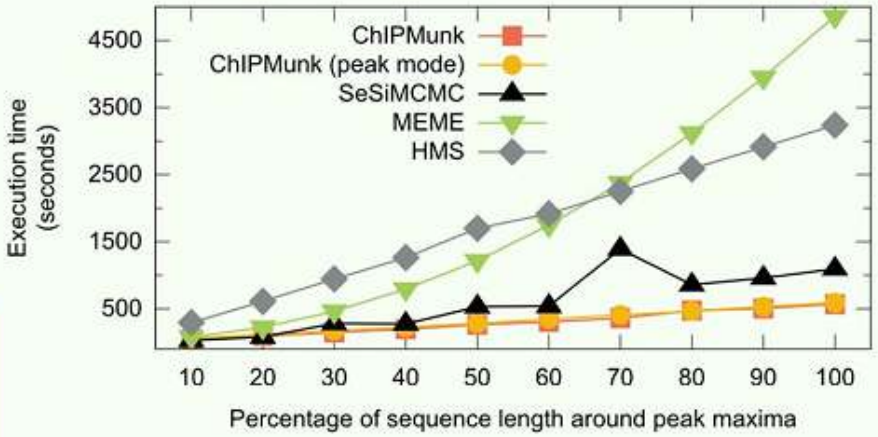
Сравнение различных программ: пример - EWS-FLI1 (Гийон, Боева, Зиновьев с соавт.)

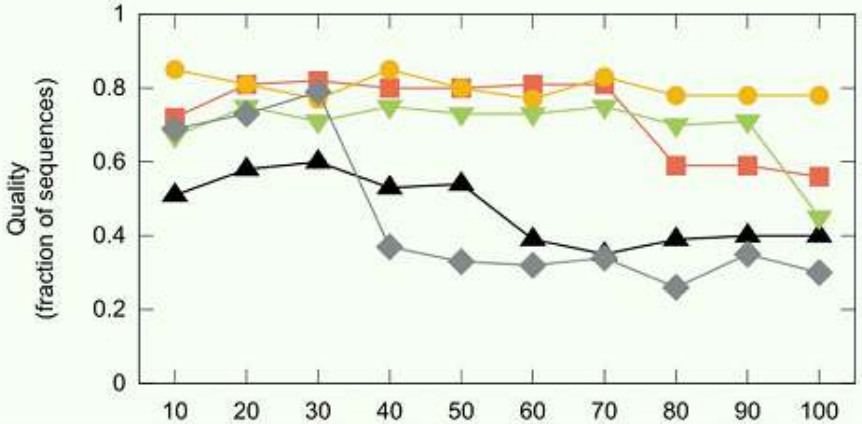
Motif discovery for EWS-FLI1

The time efficiency of motif discovery and resulting quality of the motifs identified in top 500 EWS-FLI1 peaks [Guillon et. al., 2009]

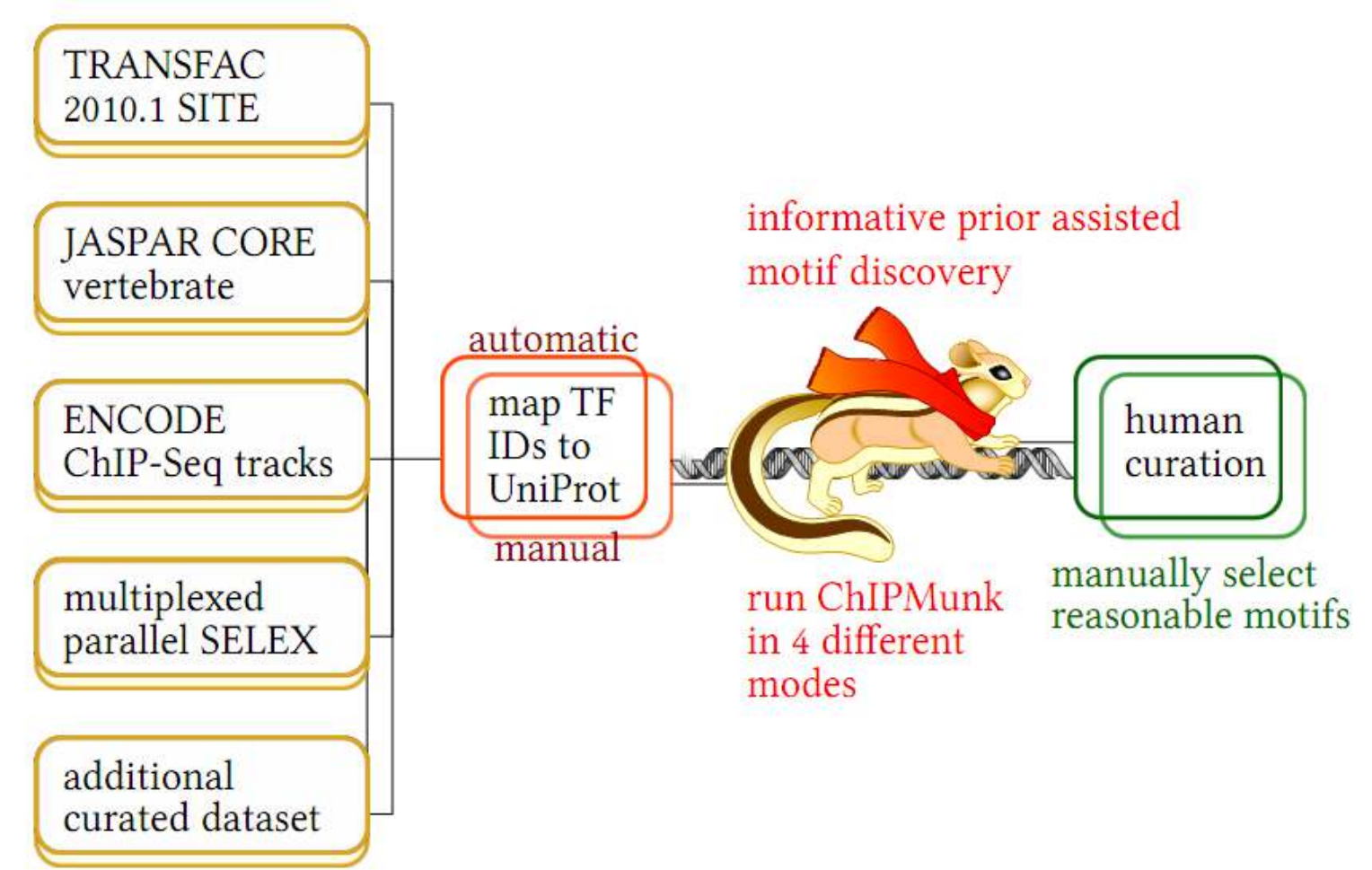
The fraction of peaks truncated to 10% of initial length centered at the peak maximum with motif hits having scores greater than the mean + 3 standard deviation of score distribution over all 11-mers was used as the measure of motif quality.





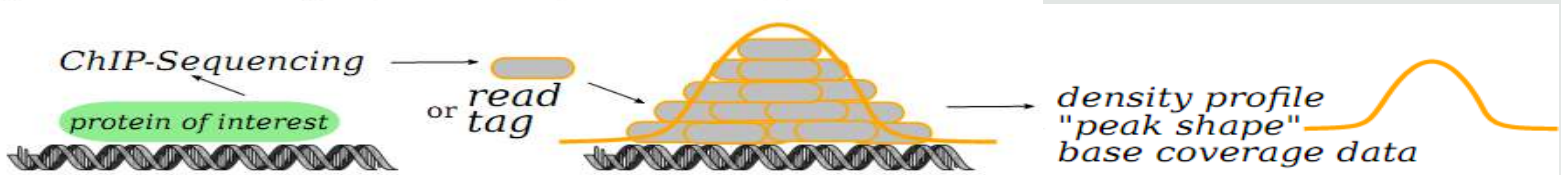


Интеграция данных, полученных различными экспериментальными методами

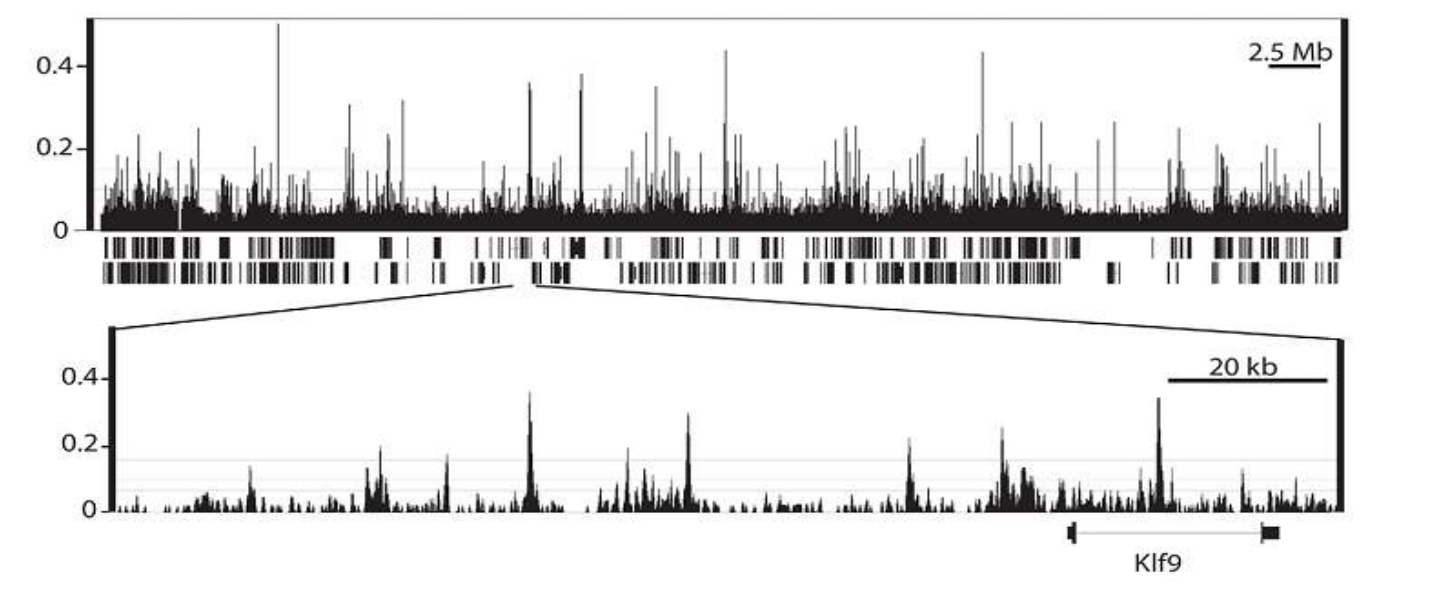


Априорная информация - профиль связывания

Use of the ChIP-Seq base coverage data:
[Hybrid Motif Sampler, Hu *et al.*, NAR 2010]



CHD7-helicase ChIP-seq, Schentz, 2010, Найти EWS_Fli



Спиральная модуляция специфического связывания

- Контакты по большой бороздке ДНК
- Существуют исключения - контакты по малой бороздке (TBP) или сложные структуры комплекса (p53) - но очень большое число белков связывается по большой бороздке

Спиральные характеристики ДНК [Schneider et al, NAR, 1997]

Определение длины мотива

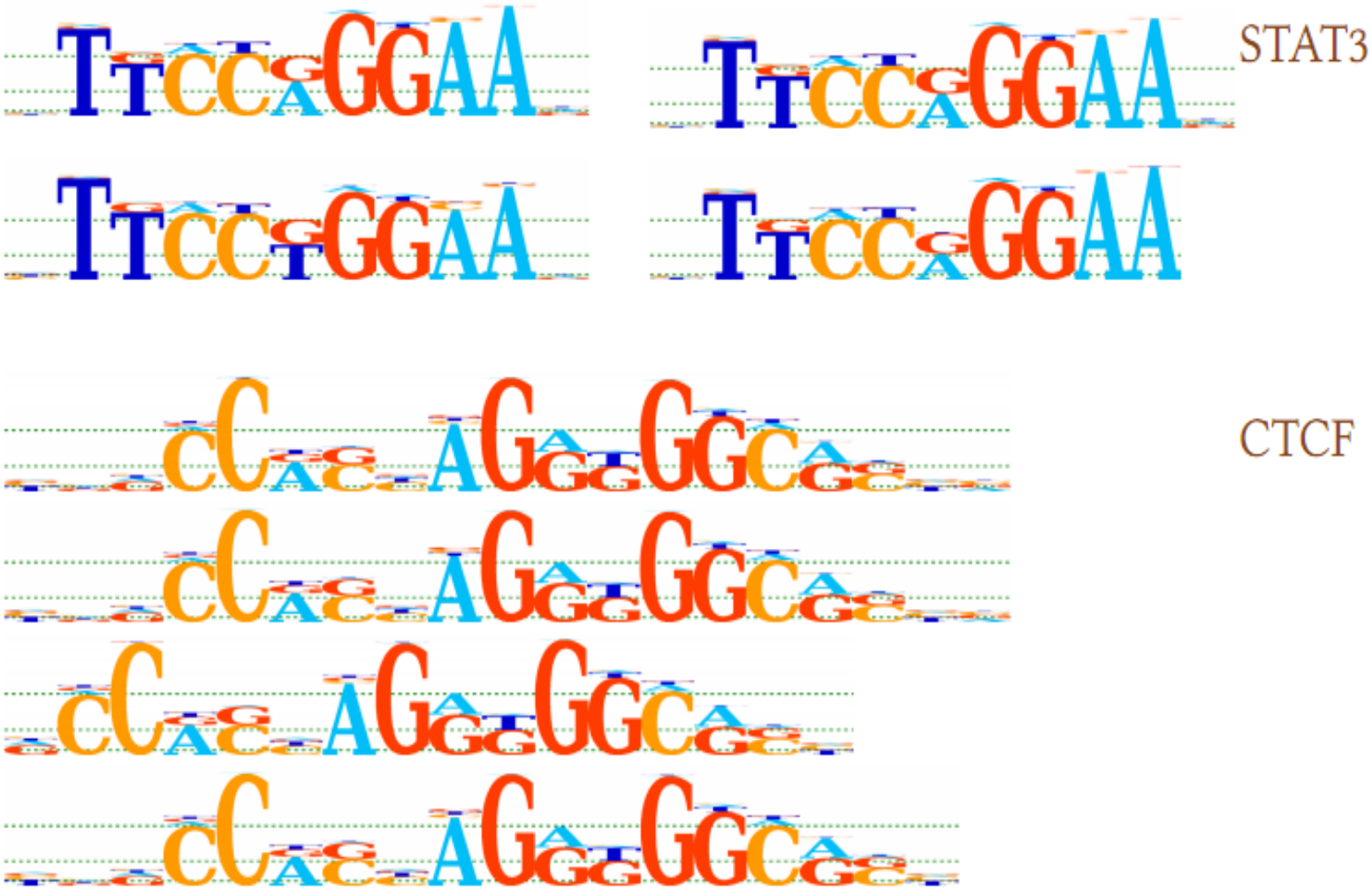
Usability and motif quality assesed in our [Kulakovskiy *et al.*, 2010, *Bioinformatics*] and independent studies [Kuttipurathu *et al*, 2011, *Bioinformatics*], [Bi *et al*, 2011, *PLoS one*];
data integration feature assesed in [Kulakovskiy *et al.*, 2009, *Biophysics*].

Четыре стратегии определения длины

- Плоский, длина увеличивается (7-25) > f2
- Плоский, длина уменьшается (25-7) > f2
- Периодический, $\cos^2(\pi^* n/T)$, T=10.5bp один > si
- Периодический, $\sin^2(\pi^* n/T)$, T=10.5bp два>do

МПВ используется как модель во всех случаях

Для многих “хороших” мотивов все модели дают сходный результат!



Homo sapiens Comprehensive Model Collection (v9 - March 2012)

[NEW] Circular tree of HOCOMOCO models (A-C quality)
Average linkage clustering strategy, distance computed by MACRO-APE at 0.0005 P-value; uniform PWM normalization; picture by jsPhyloSVG.

Details

HOCOMOCO-AD curated: high-confindence models

The 'A to D'-part of the curated collection contains **manually curated** TFBS models.

We have used the high-performance motif discovery tool ChIPMunk to produce TFBS models for human TFs integrating data from different sources.
ChIPMunk was used in 4 different modes searching for optimal models in the lengths range from 7 to 22bp.
Then we manually selected the most reasonable models for each transcription factor and assigned the quality ratings from 'A' (best) to 'F' (fail).
The collection of **'A' to 'D'** quality models represents the most reasonable subset to be used for further analysis.

HOCOMOCO-E curated: low-confidence models

The 'E'-part of the curated collection contains low confidence models.

HOCOMOCO-FULL: full collection

The whole collection listing 4 models (f1,f2,si,do) for each TF.

Downloads

HOCOMOCO-AD

download the AD **curated** collection, small-BiSMark XML format, **uniform** background for PWMs, one-file-per-model

download the AD **curated** collection, small-BiSMark XML format, **hg19** background for PWMs, one-file-per-model

download the AD **curated** collection, plain text format, **weighted** position count matrices (WPCM), rows as letters (ACGT)

http://autosome.ru/HOCOMOCO/addon/ave_clustered_ave_displayed

home HOCOMOCO-AD build9-MAR2012

Click the 'motif name' to see detailed information for the selected motif.

motif	UniProt	quality	LOGO	weight
AHR_si	AHR	B		153.96
AIRE_f2	AIRE	C		41.31
ALX1_si	ALX1	B		31.0
ANDR_do	ANDR	B		130.0
AP2A_f2	AP2A	A		1929.74
AP2B_f1	AP2B	B		27.91
AP2C_f1	AP2C	A		1599.02
AP2D_f1	AP2D	D		5.0
ARI3A_do	ARI3A	D		44.15
ARI3A_f1	ARI3A	D		47.11
ARI5B_f1	ARI5B	C		15.0
ARNT2_si	ARNT2	D		18.0
ARNT_f1	ARNT	B		267.8
ATF1_si	ATF1	B		33.75
ATF2+ATF4_f1	ATF2 ATF4	B		89.79
ATF3_f1	ATF3	A		2036.0
ATF5_si	ATF5	C		33.0
ATF6A_si	ATF6A	B		20.0
BACH1_si	BACH1	C		8.0
BARX2_si	BARX2	D		14.0
BATF_si	BATF	B		1878.11
BCL6_f1	BCL6	C		32.64
BHE40_f2	BHE40	A		2037.0
BHE41_f1	BHE41	D		10

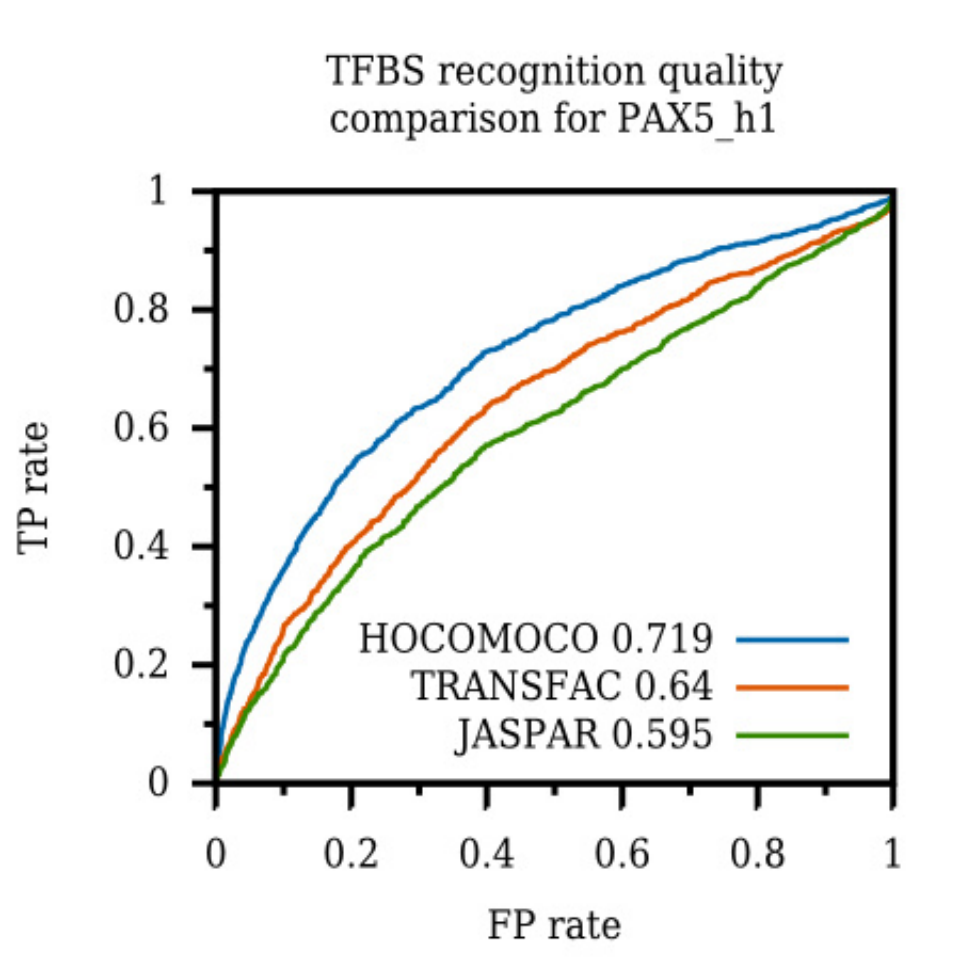
http://www.uniprot.org/uniprot/ATF5_H

Общая статистика (на март 2012)

Ка-че-ство	Число ТФ	Число Моделей	Медиана – Данные/ТФ	Медиана – число источников данных на ТФ
A	49	52	2037	2
B	82	87	159.5	2
C	128	139	38.0	1
D	142	148	16.5	1
E	55	55	11.0	1
F	18	1	2047.0	1

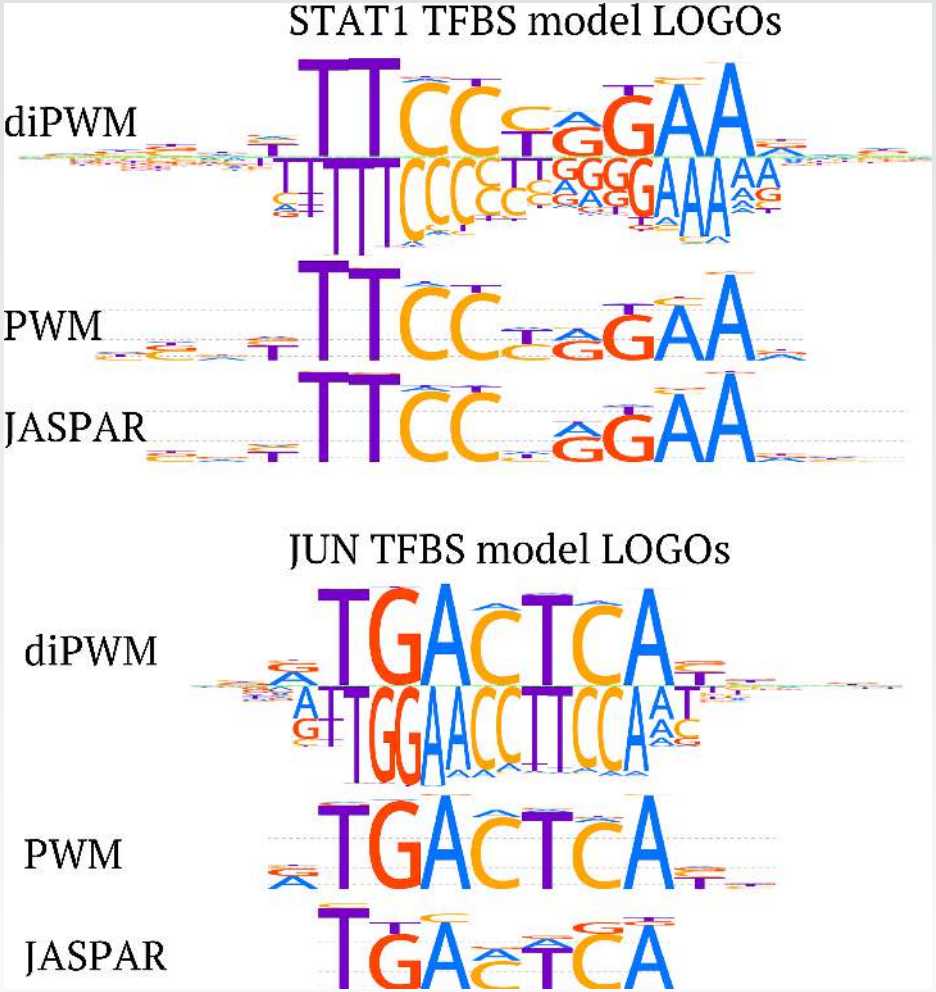
ROC - кривая

- 1000 «слабых» пиков ChIP-seq - не использованных в построении мотива

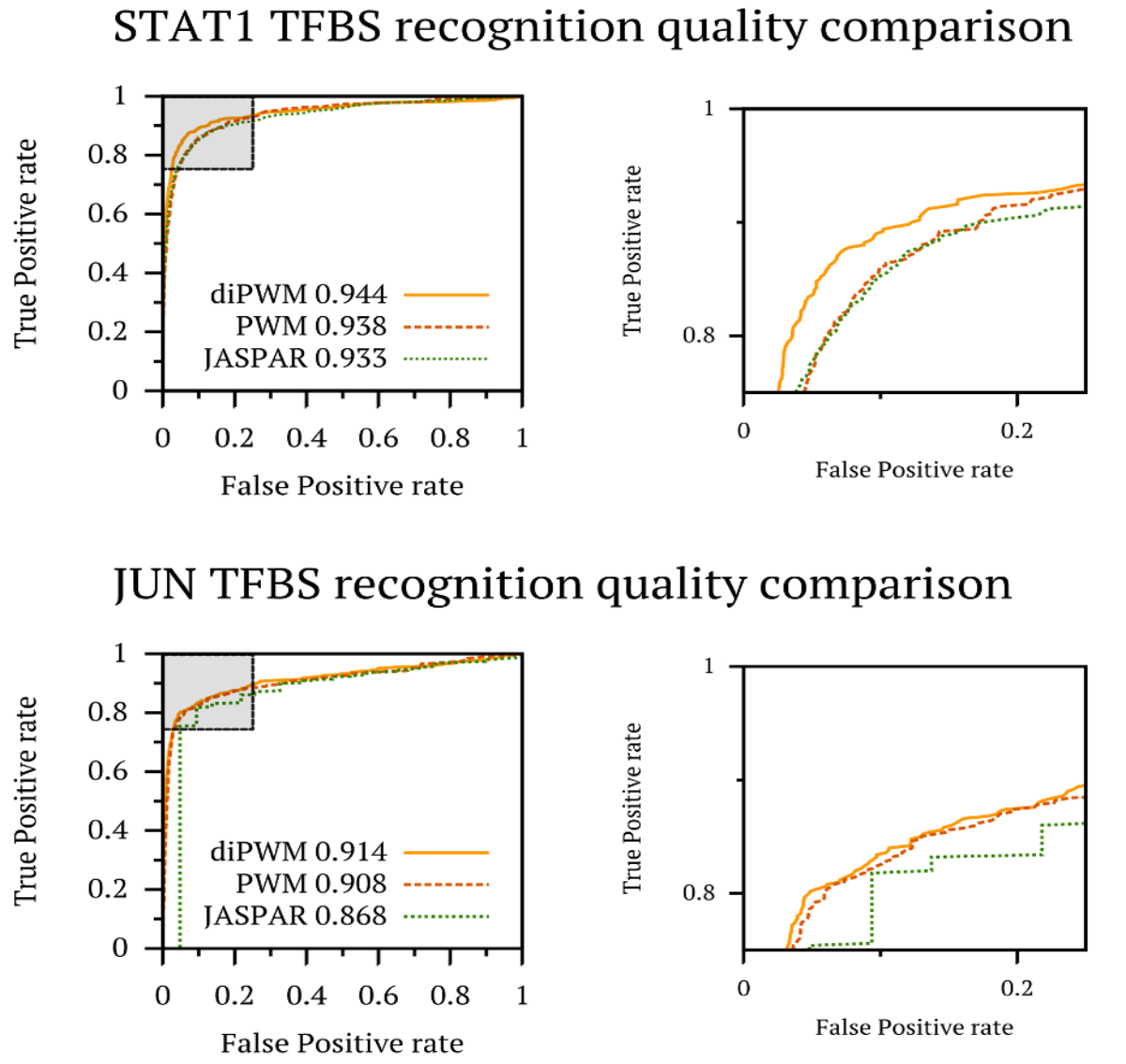


- TP-rate – сайт найден в пике
- FP-rate – теоретическая вероятность нахождения сайта в случайной последовательности, с длиной, равной средней ширине пика

Матрицы пар нуклеотидов



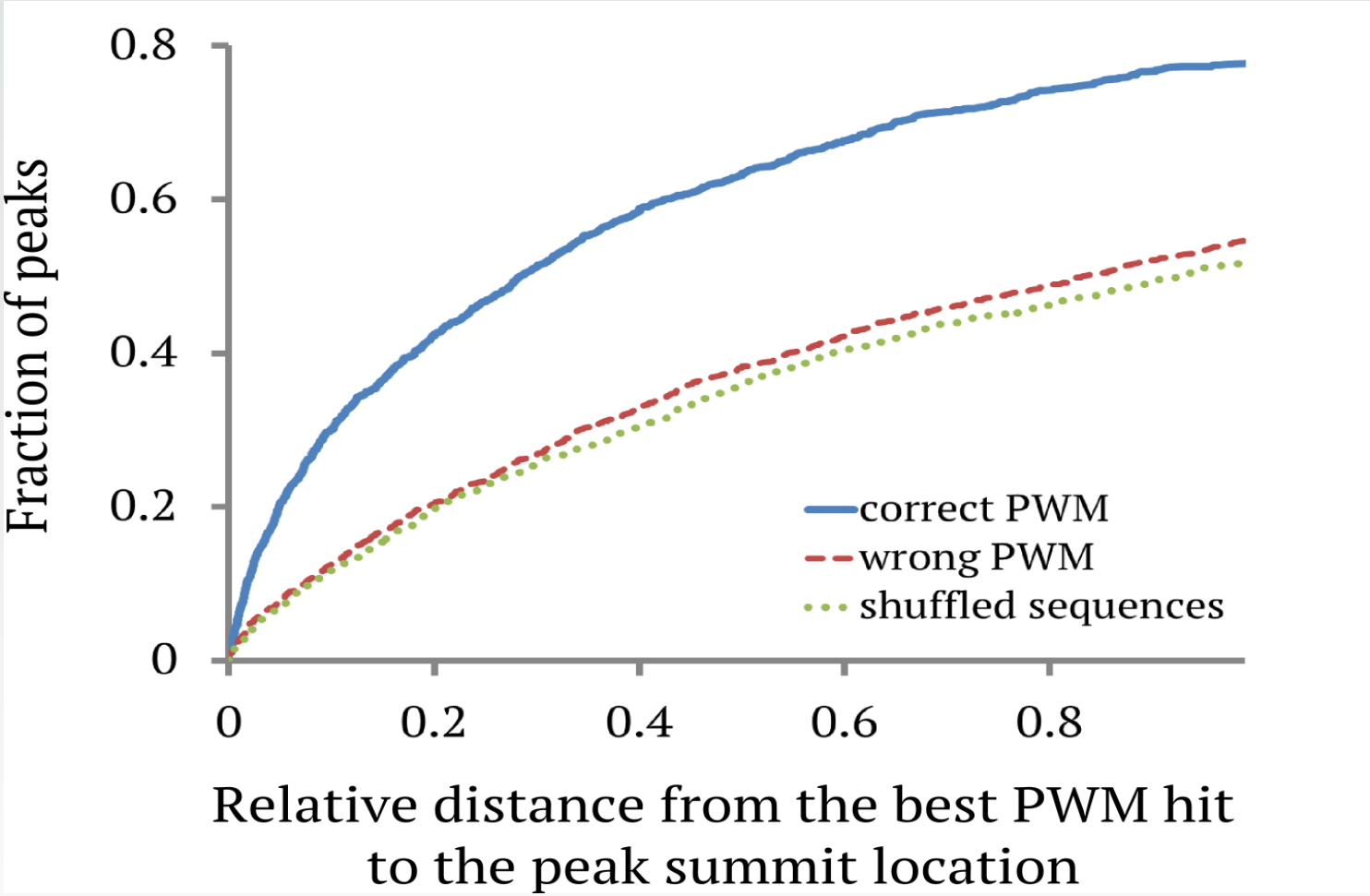
Иногда дают выигрыш – зависит от ТФ



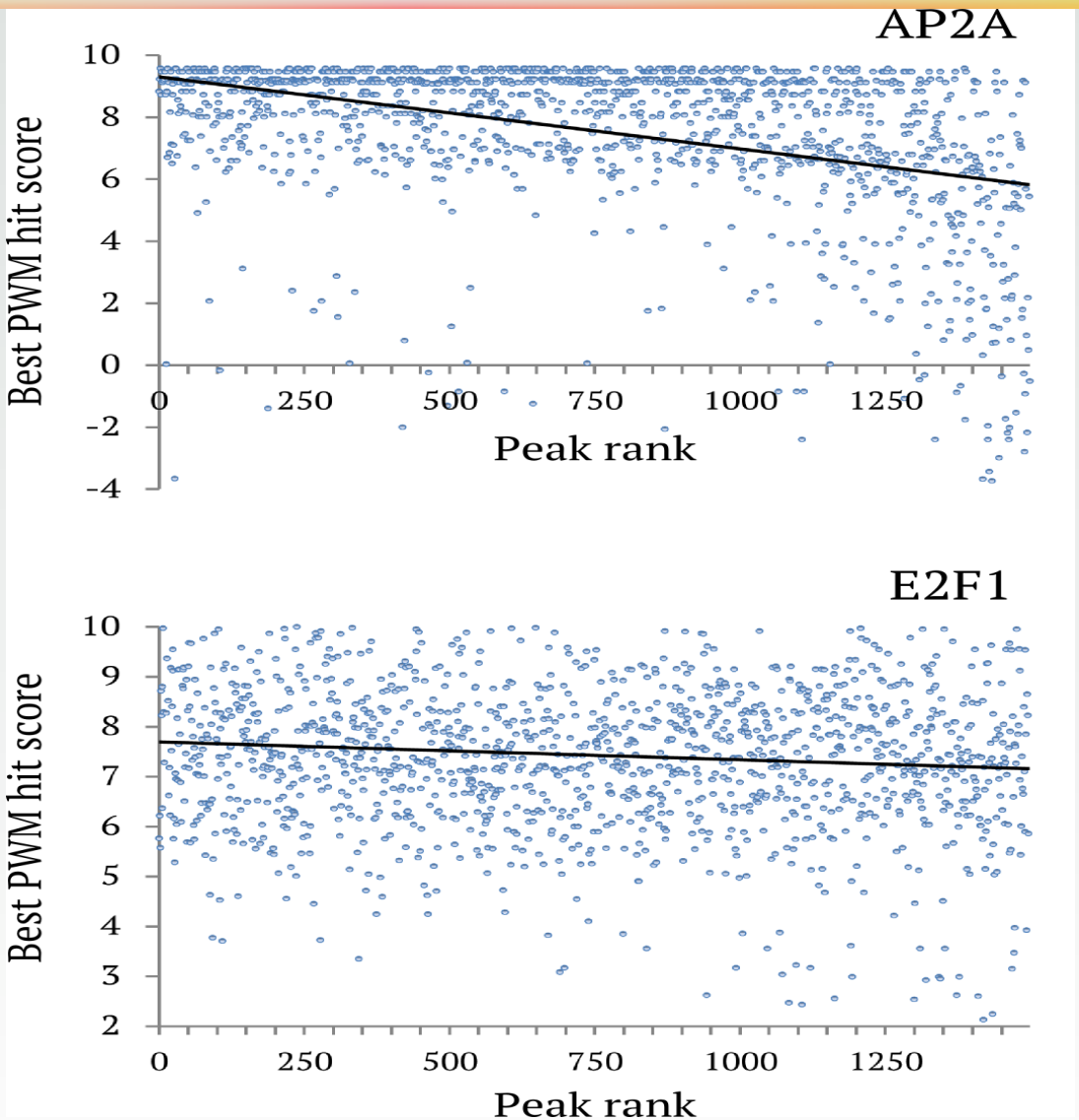
Роль огибающей

- Из 401 модели
- - 114 одиночный бокс
- - 26 двойной бокс

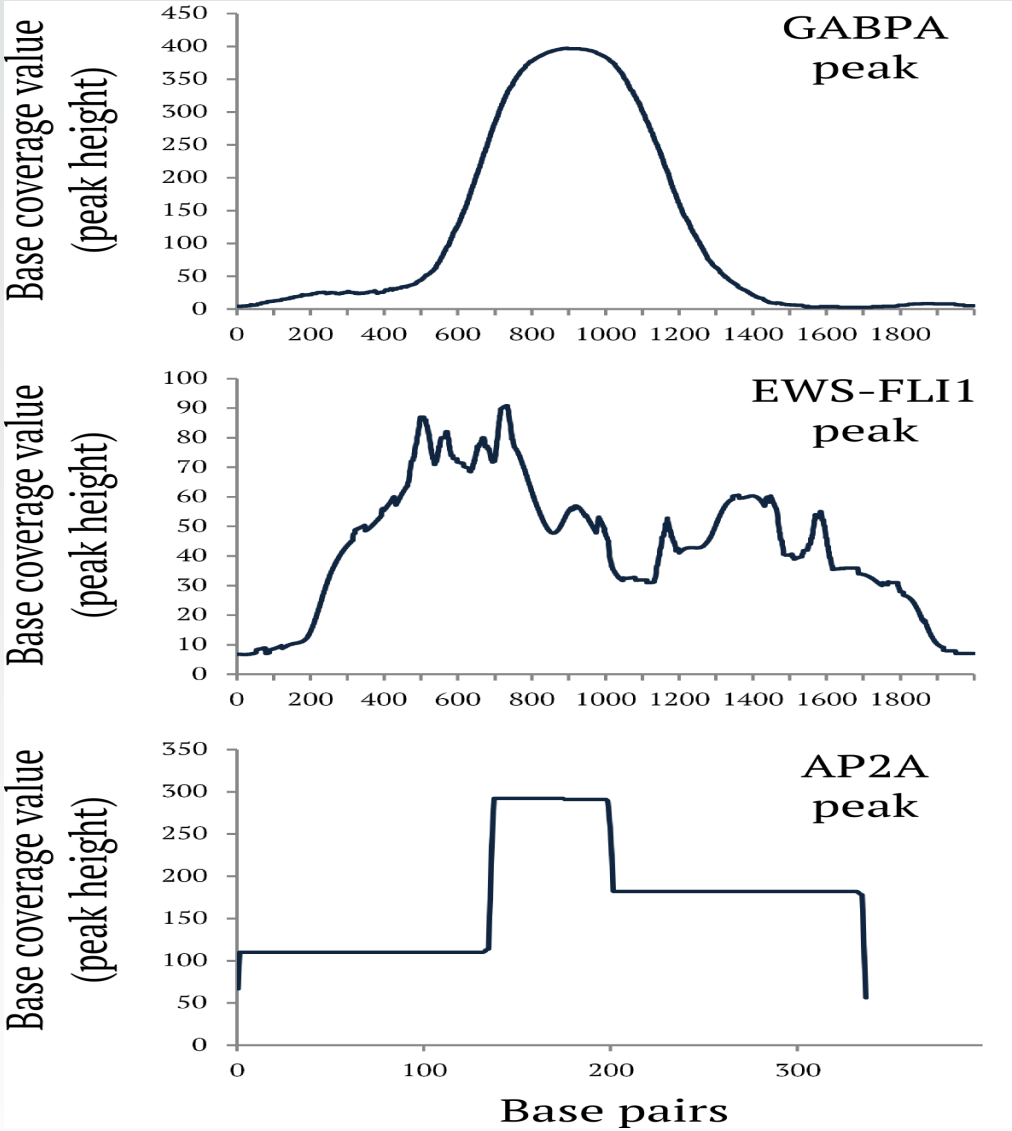
Факторы действительно связываются в пиках!



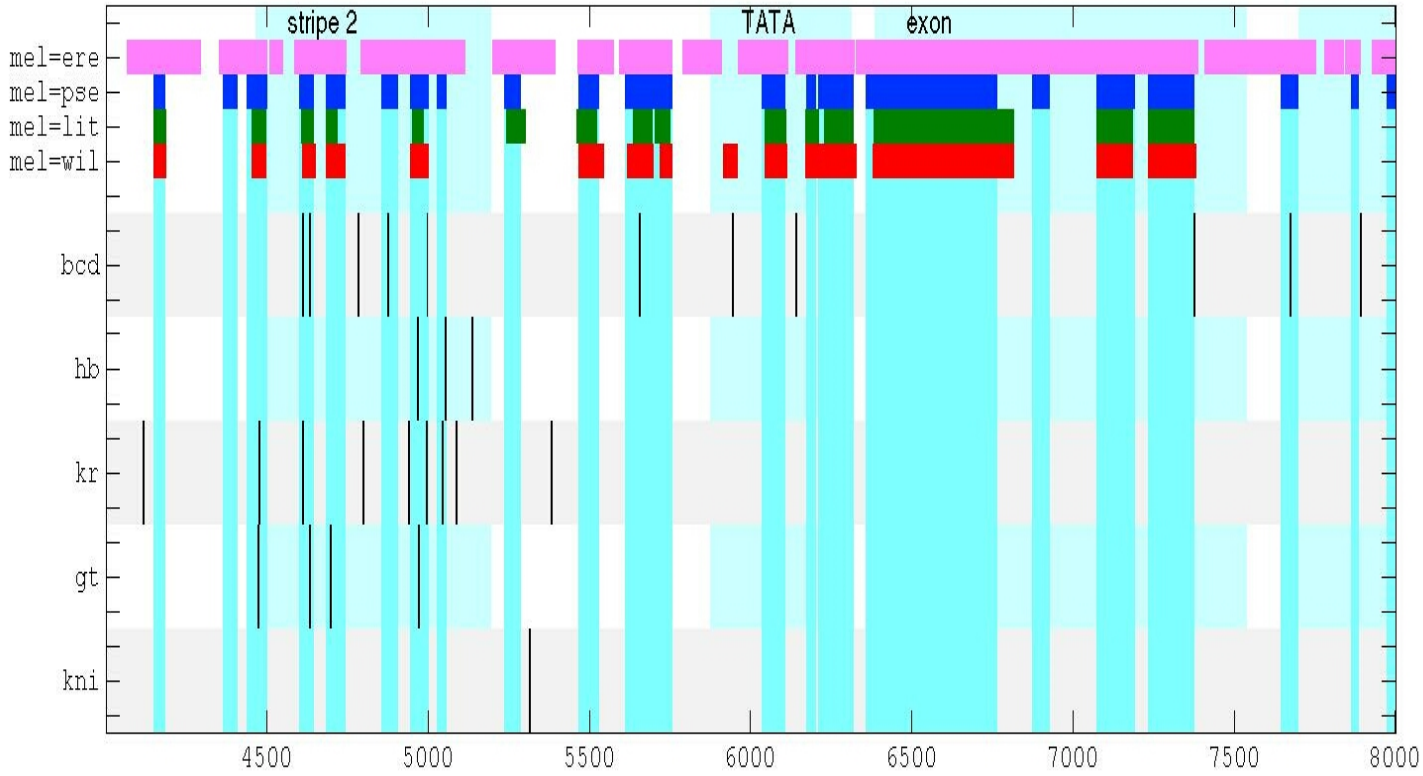
Более высокие пики имеют больший скор матрицы



Пики имеют разную форму

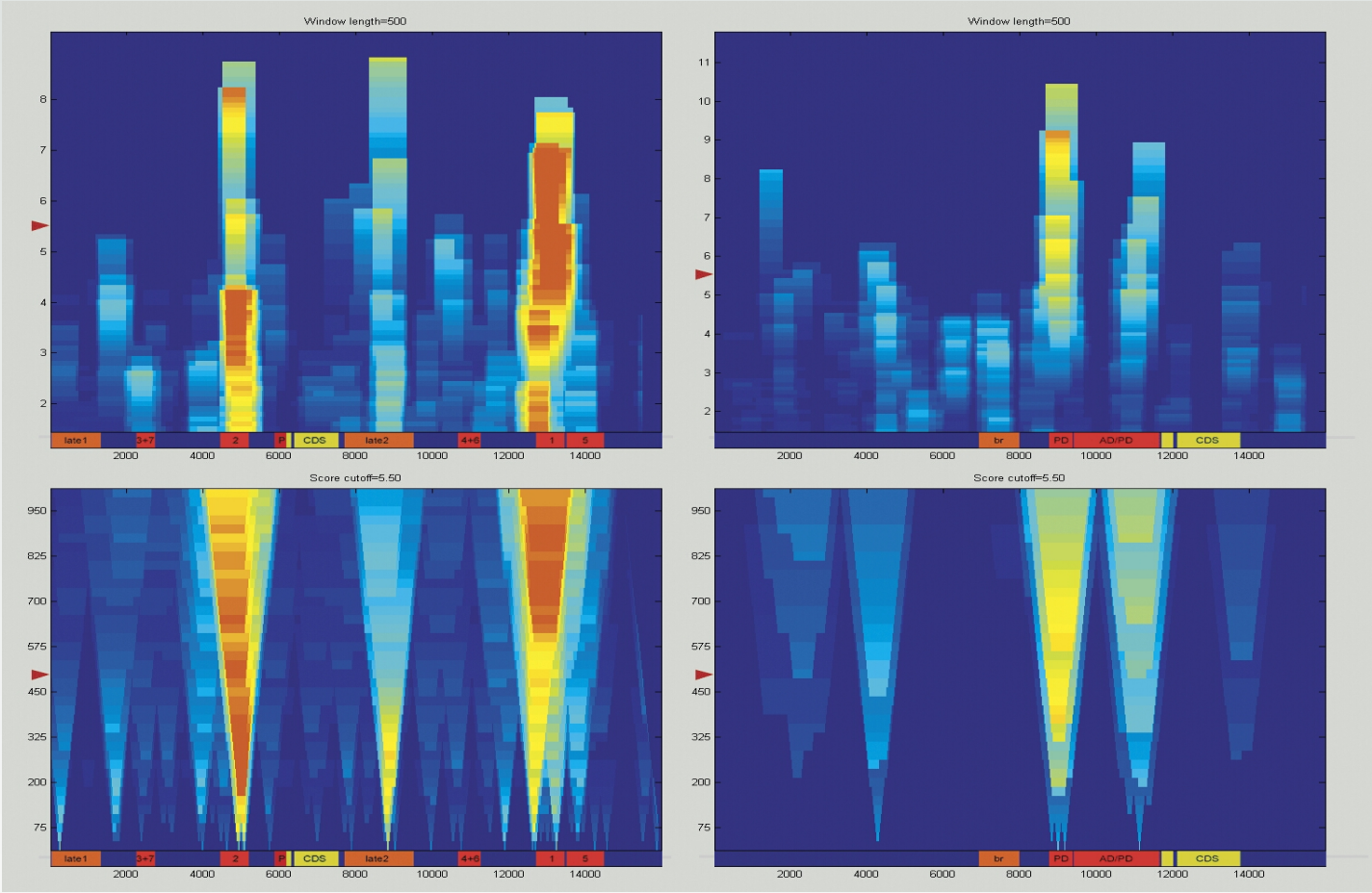


Кластеры сайтов связывания



Регуляторные модули часто содержат несколько сайтов связывания одного и того же фактора

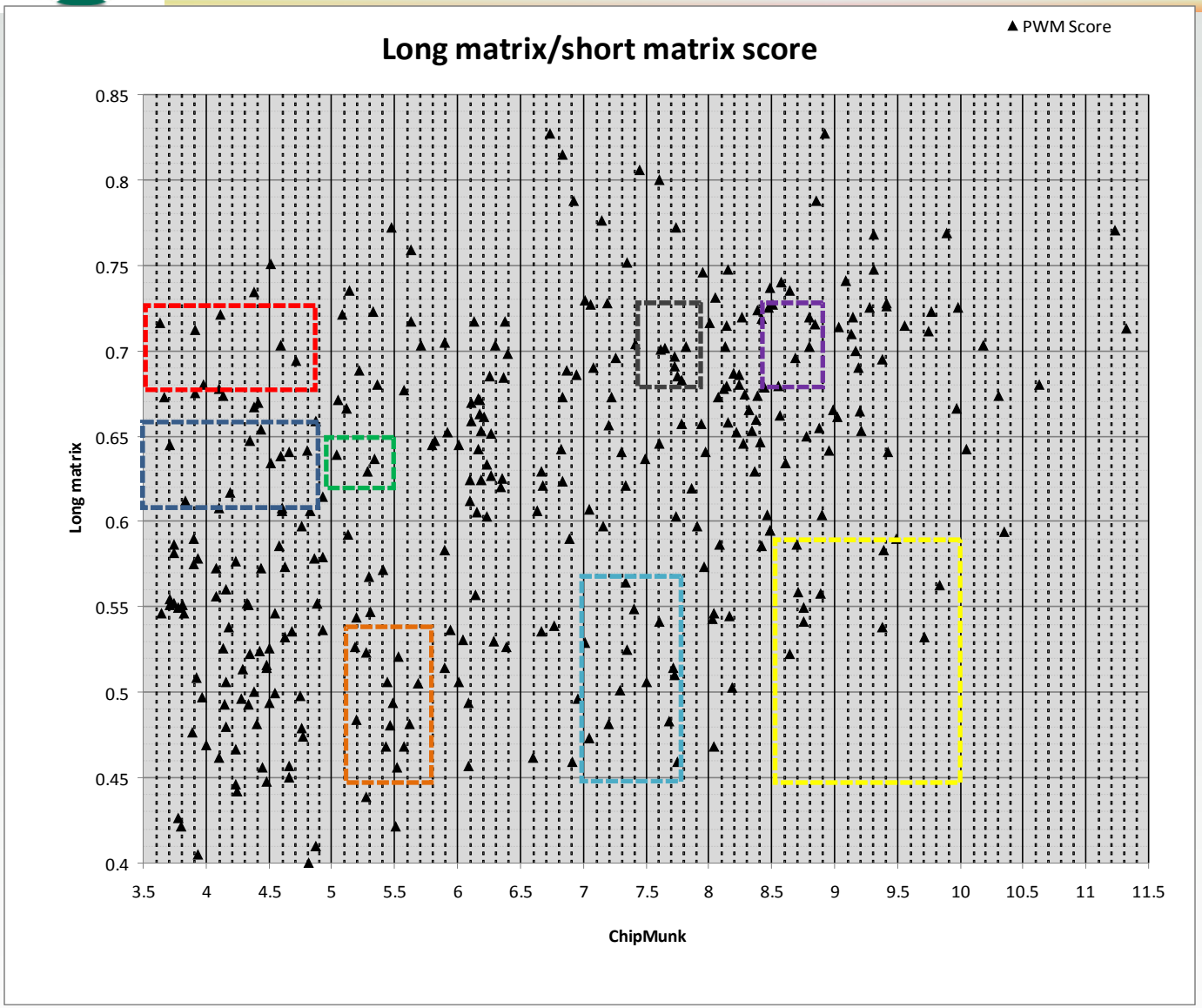
Могут быть использованы для поиска регуляторных элементов



Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA.. Genome Res. 2003 Apr;13(4):579-88.

Markstein M, Markstein P, Markstein V, Levine MS. Proc Natl Acad Sci U S A. 2002 Jan 22;99(2):763-8.

И объяснять наличие пиков с очень слабыми сайтами

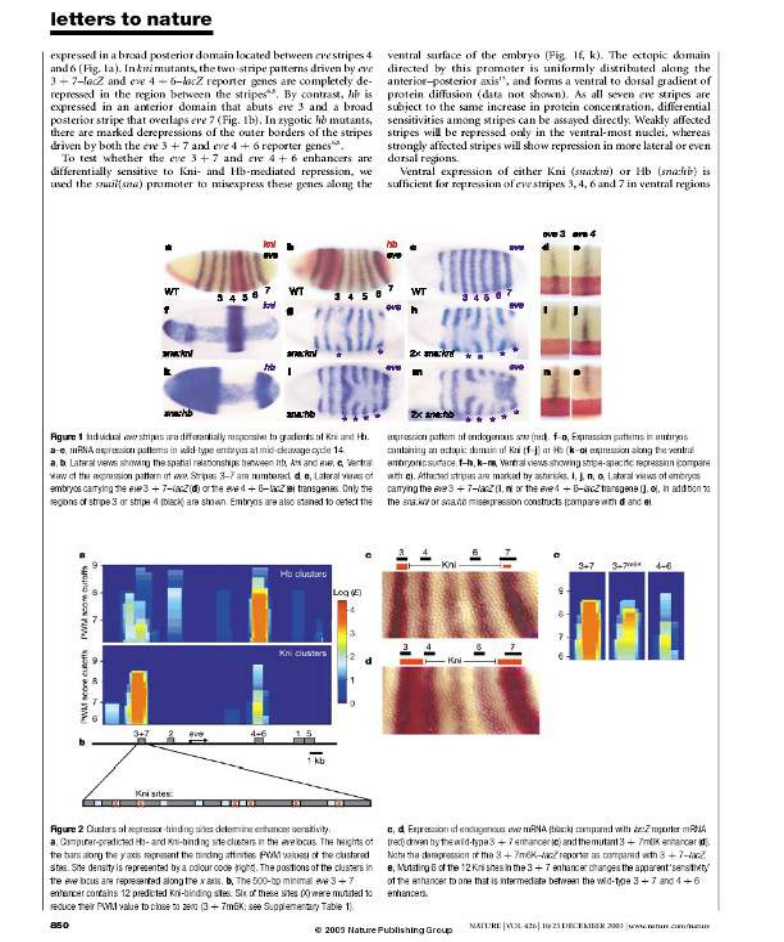


Merkulova TI,

Oschepkov DA,

Levitzkiy VG

Связывание кооперативное



letters to nature

expressed in a broad posterior domain located between *eve* stripes 4 and 6 (Fig. 1a). In *kni* mutants, the two-stripe patterns driven by *eve* 3 + 7–*lacZ* and *eve* 4 + 6–*lacZ* reporter genes are completely derepressed in the region between the stripes[6,8]. By contrast, *hb* is expressed in an anterior domain that abuts *eve* 3 and a broad posterior stripe that overlaps *eve* 7 (Fig. 1b). In zygotic *hb* mutants, there are marked derepressions of the outer borders of the stripes driven by both the *eve* 3 + 7 and *eve* 4 + 6 reporter genes[6,8].

To test whether the *eve* 3 + 7 and *eve* 4 + 6 enhancers are differentially sensitive to Kni- and Hb-mediated repression, we used the *snail*(*sna*) promoter to misexpress these genes along the ventral surface of the embryo (Fig. 1f, k). The ectopic domain directed by this promoter is uniformly distributed along the anterior–posterior axis[14], and forms a ventral to dorsal gradient of protein diffusion (data not shown). As all seven *eve* stripes are subject to the same increase in protein concentration, differential sensitivities among stripes can be assayed directly. Weakly affected stripes will be repressed only in the ventral-most nuclei, whereas strongly affected stripes will show repression in more lateral or even dorsal regions.

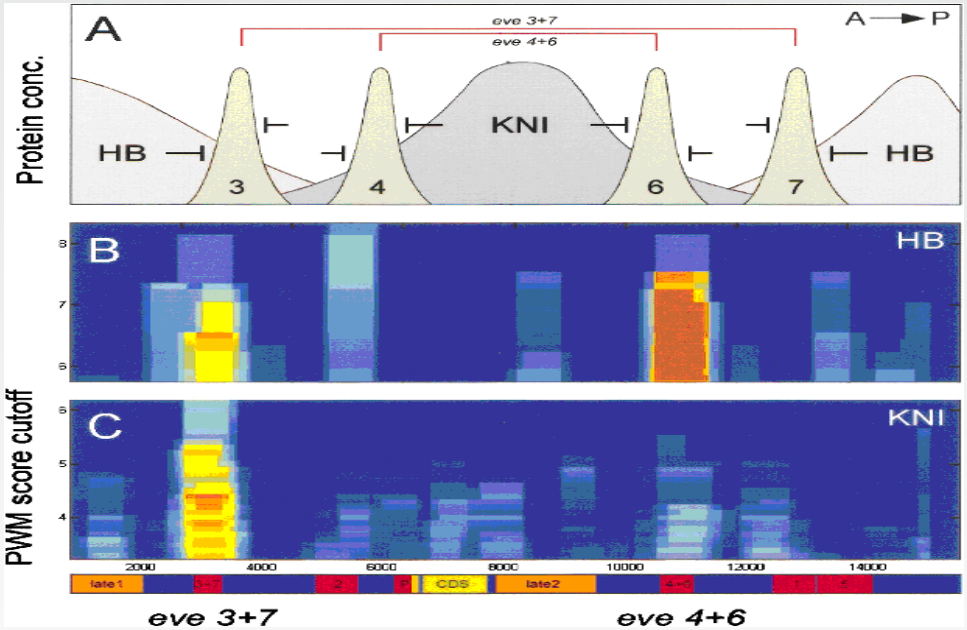
Ventral expression of either Kni (*sna:kni*) or Hb (*sna:hb*) is sufficient for repression of *eve* stripes 3, 4, 6 and 7 in ventral regions

Figure 1 Individual *eve* stripes are differentially responsive to gradients of Kni and Hb. **a**–**e**, mRNA expression patterns in wild-type embryos at mid-cleavage cycle 14. **a**, **b**, Lateral views showing the spatial relationships between *hb*, *kni* and *eve*. **c**, Ventral view of the expression pattern of *eve*. Stripes 3–7 are numbered. **d**, **e**, Lateral views of embryos carrying the *eve*3 + 7–*lacZ* (**d**) or the *eve*4 + 6–*lacZ* (**e**) transgenes. Only the regions of stripe 3 or stripe 4 (black) are shown. Embryos are also stained to detect the expression pattern of endogenous *eve* (red). **f**–**o**, Expression patterns in embryos containing an ectopic domain of Kni (**f**–**j**) or Hb (**k**–**o**) expression along the ventral embryonic surface. **f**–**h**, **k**–**m**, Ventral views showing stripe-specific repression (compare with **c**). Affected stripes are marked by asterisks. **i**, **j**, **n**, **o**, Lateral views of embryos carrying the *eve*3 + 7–*lacZ* (**i**, **n**) or the *eve*4 + 6–*lacZ* transgene (**j**, **o**), in addition to the *sna:kni* or *sna:hb* misexpression constructs (compare with **d** and **e**).

Figure 2 Clusters of repressor-binding sites determine enhancer sensitivity. **a**, Computer-predicted Hb- and Kni-binding site clusters in the *eve* locus. The heights of the bars along the *y* axis represent the binding affinities (PWM values) of the clustered sites. Site density is represented by a colour code (right). The positions of the clusters in the *eve* locus are represented along the *x* axis. **b**, The 500-bp minimal *eve* 3 + 7 enhancer contains 12 predicted Kni-binding sites. Six of these sites (X) were mutated to reduce their PWM value to close to zero (3 + 7mBK; see Supplementary Table 1). **c**, **d**, Expression of endogenous *eve* mRNA (black) compared with *lacZ* reporter mRNA (red) driven by the wild-type 3 + 7 enhancer (**c**) and the mutant 3 + 7mBK enhancer (**d**). Note the derepression of the 3 + 7mBK–*lacZ* reporter as compared with 3 + 7–*lacZ*. **e**, Mutating 6 of the 12 Kni sites in the 3 + 7 enhancer changes the apparent 'sensitivity' of the enhancer to one that is intermediate between the wild-type 3 + 7 and 4 + 6 enhancers.

850 © 2003 Nature Publishing Group NATURE | VOL 426 | 18/25 DECEMBER 2003 | www.nature.com/nature

… to provide for on/off regulation of expression response (by a steeper concentration-response curve)



D.E. Clyde, M.|S.G. Colorado, X. Wu, A. Paré, D. Papatsenko, S. Small. (2003) *Nature*, vol. 426,, pp. 849-853

… и «аддитивное»

Provide for smooth regulation of expression response (by averaging contribution of several independently bound factors)

Cell PRESS

Molecular Cell
Article

Noncooperative Interactions between Transcription Factors and Clustered DNA Binding Sites Enable Graded Transcriptional Responses to Environmental Inputs

Luca Giorgetti,[1] Trevor Siggers,[2] Guido Tiana,[3,4] Greta Caprara,[1] Samuele Notarbartolo,[1] Teresa Corona,[6] Manolis Pasparakis,[4] Paolo Milani,[4,5] Martha L. Bulyk,[2,7,8] and Gioacchino Natoli[1,*]
[1]Department of Experimental Oncology, European Institute of Oncology, IFOM-IEO Campus, 20139 Milan, Italy
[2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[3]INFN
[4]Dipartimento di Fisica
[5]CIMAINA
Universita' di Milano, 20133 Milan, Italy
[6]Institute for Genetics, University of Cologne, 50674 Cologne, Germany
[7]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[8]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA
*Correspondence: gioacchino.natoli@ifom-ieo-campus.it
DOI 10.1016/j.molcel.2010.01.016

SUMMARY

A paradigm in transcriptional regulation is that graded increases in transcription factor (TF) concentration are translated into on/off transcriptional responses by cooperative TF binding to adjacent sites. Digital transcriptional responses underlie the definition of anatomical boundaries during development. Here we show that NF-κB, a TF controlling inflammation and immunity, is conversely an analog transcriptional regulator that uses clustered binding sites noncoop-

scriptional response, enabled by cooperative TF binding to adjacent sites (Segal et al., 2008) (Figure 1). This paradigm applies well to TFs operating in developmental processes, notably embryonic segmentation, which require the definition of sharp anatomical borders. A classic example is the morphogen model of Bicoid (Bcd) action in *Drosophila* embryogenesis. Cooperative binding of Bcd to adjacent DNA sites represents the molecular basis by which a gradual protein gradient of Bcd along the anterior-posterior embryonic axis is converted into a sharp on/off pattern of expression of Bcd target genes (Burz et al., 1998; Gregor et al., 2007). Nonetheless, clusters of adjacent binding sites

Molecular Cell

Graded Environmental Transcriptional Responses

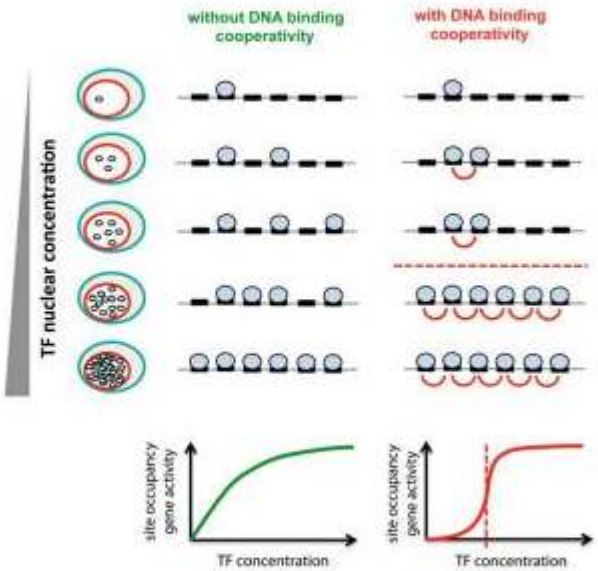


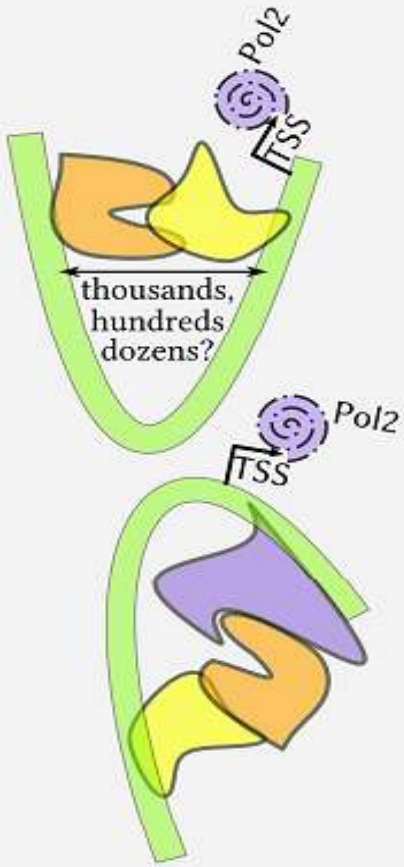
Figure 1. DNA Binding Cooperativity Generates Sharp Transcriptional Boundaries

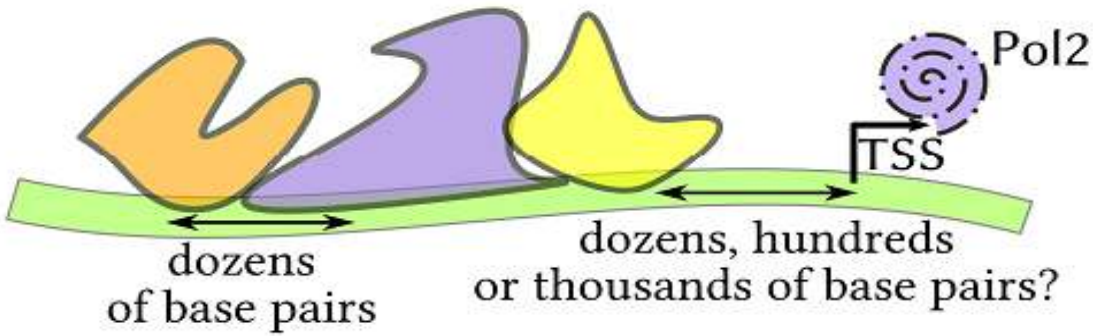
If a TF binds to a cluster of DNA binding sites in a noncooperative manner, a gradual increase in TF concentration generates a gradual increase in the average occupancy of the cluster. Conversely, if TF binding to adjacent sites is cooperative, a gradual increase in TF concentration generates a "digital" on/off response as the concentration sweeps a threshold value (dashed line). The higher binding cooperativity, the steeper the transition between the "off" and the "on" states.

Расстояние между гетеросайтами и структура инициаторного комплекса. Композитные элементы

Transcription Factor Binding Sites

and composite elements in regulatory regions



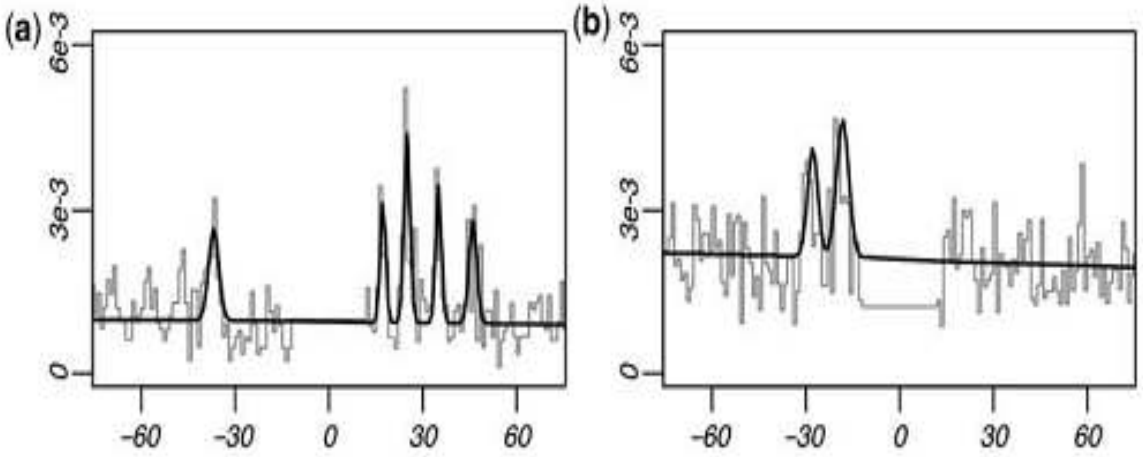


typical ideas:

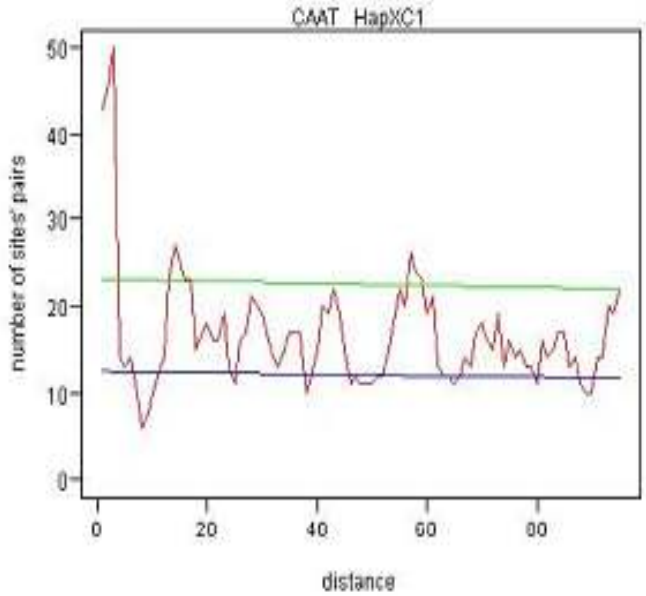
- TFs can form different complexes (combinatorial regulation);
- mutual orientation of TF binding sites is important;
- detectable and effective TFBS distance preferences ~dozens of base pairs.

The TFBS distance preferences scale problem for different TF configurations.

Опубликованные картины предпочтительных расстояний



Yokoyama et al., *NAR* **2009**
Distribution of distances between NFY-NFY (a) and NFY-SP1 (b) binding site pairs. NFY site at zero. IUPAC consensus as a motif model.

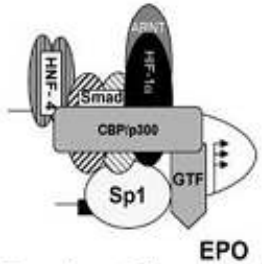


Shelest et al., *Bioinformatics* **2010**
Distribution of distances between CAAT and HapXC1 binding sites. Small datasets used, no information on mutual orientation.

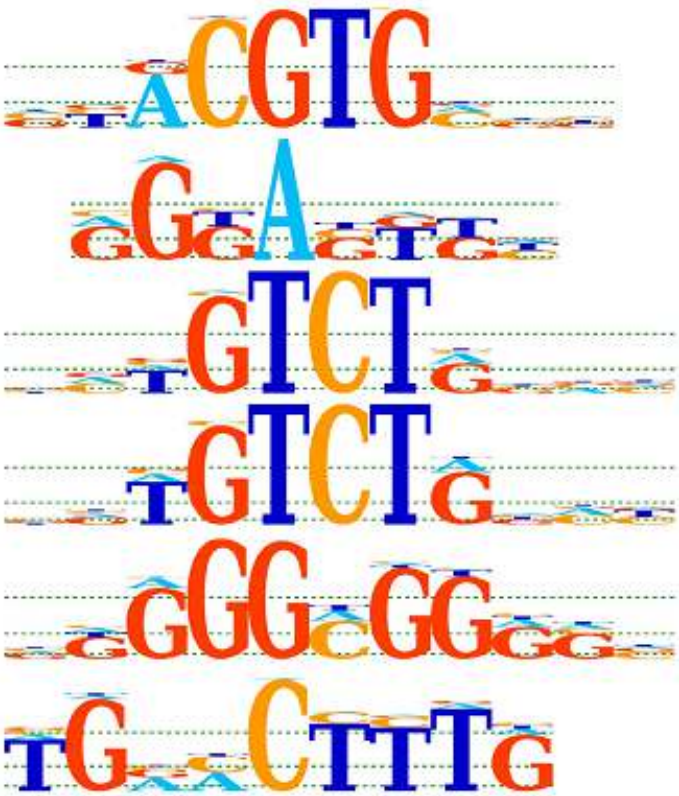
Other examples:
Mercier et al., *PLoS ONE* 2011;
Laurila et al., *NAR* 2009.

Пример: HIF-1a, на 3’ энхансере эритропоэтина

Regulation of erithropoetin (EPO) gene expression in hypoxia response in human cells: a case study



Sanchez-Elsner et al., *J Mol Biol* 2004



HRE (hypoxia-responsive element, HIF1A-ARNT binding site)

EP300

SMAD4

SMAD3

SP1

HNF4

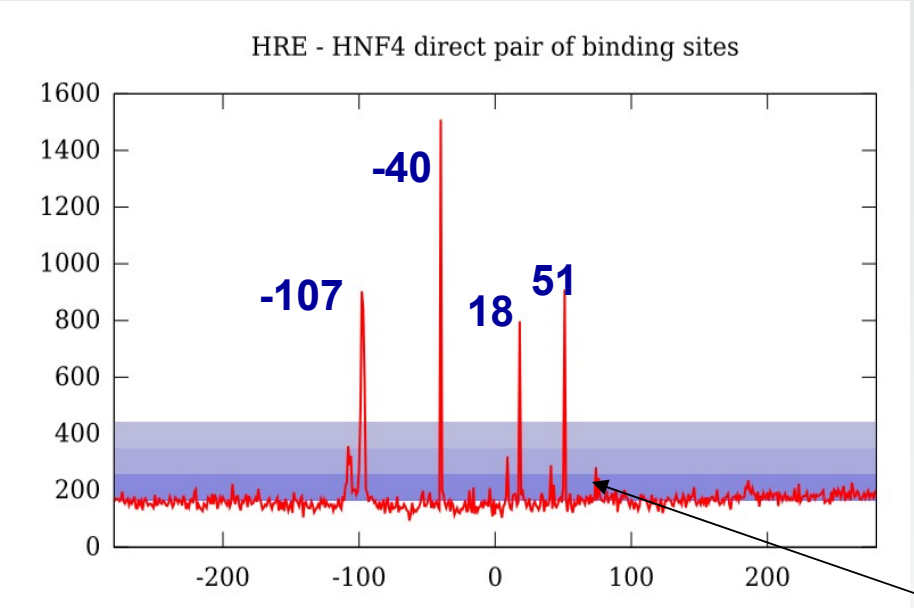
TRANSFAC data, Motif discovery by ChIPMunk

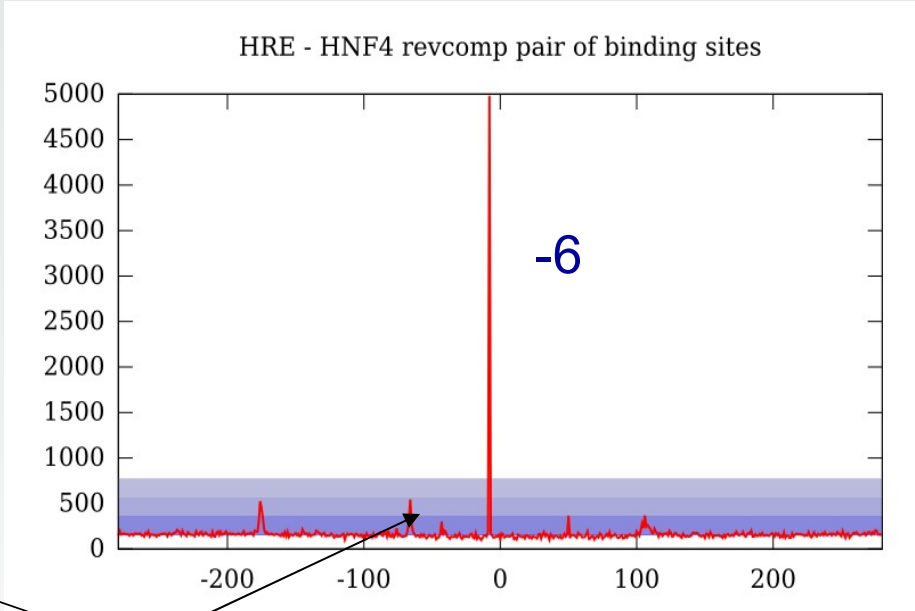
http://line.imb.ac.ru/ChIPMunk

Motif logos for PWMs used in the study.

Профили предпочтения расстояний (HIF1a - HNF4 пример)

Косвенное подтверждение качества мотивов:
– сильный рост качества позиционных предпочтений





HRE элемент (сайт HIF1a) расположен в нуле

Сайты, расположенные в разных ориентациях имеют различные позиционные предпочтения

Kulakovskiy et al, Bioinformatics, (2011), vol. 27, pp 2621

Определение первичных мишеней гипоксии

Мишени HIF1a в геноме человека

- Предсказано до 3000 мишеней, включая многие (VEGF), не выявленные другими методами

Процент истинных предсказаний на известной выборке 120 генов ~ 0.6 (74 гена)

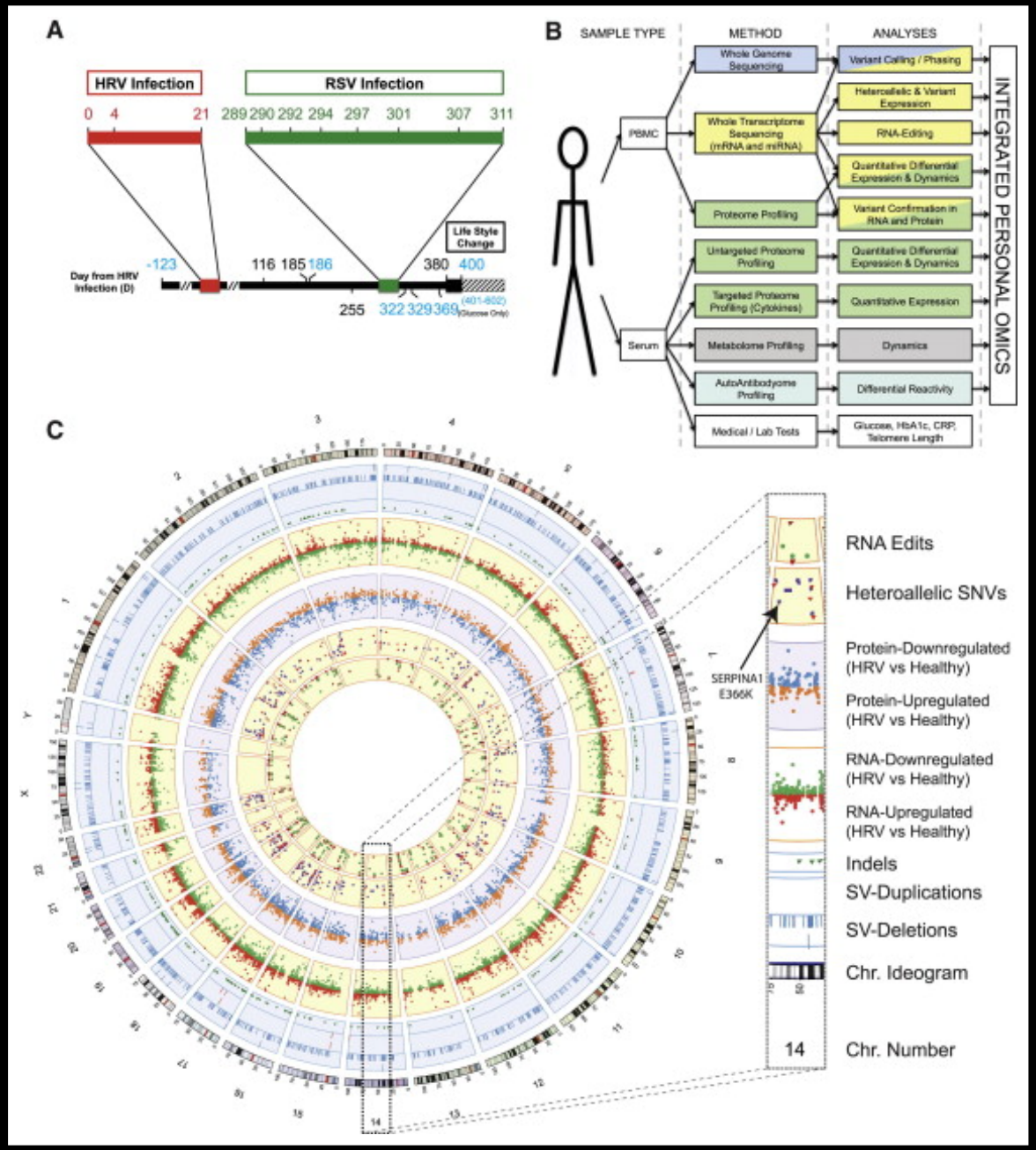
Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes

Rui Chen, George I. Mias, Jennifer Li-Pook-Than, Lihua Jiang, Hugo Y.K. Lam, Rong Chen, Elana Miriami, Konrad J. Karczewski, Manoj Hariharan, Frederick E. Dewey, Yong Cheng, Michael J. Clark, Hogune Im, Lukas Habegger, Suganthi Balasubramanian, Maeve O'Huallachain, Joel T. Dudley, Sara Hillenmeyer, Rajini Haraksingh, Donald Sharon, Ghia Euskirchen, Phil Lacroute, Keith Bettinger, Alan P. Boyle, Maya Kasowski, Fabian Grubert, Scott Seki, Marco Garcia, Michelle Whirl-Carrillo, Mercedes Gallardo, Maria A. Blasco, Peter L. Greenberg, Phyllis Snyder, Teri E. Klein, Russ B. Altman, Atul J. Butte, Euan A. Ashley, Mark Gerstein, Kari C. Nadeau, Hua Tang and Michael Snyder

Cell
Volume 148, Issue 6, Pages 1293-1307 (March 2012)
DOI: 10.1016/j.cell.2012.02.009

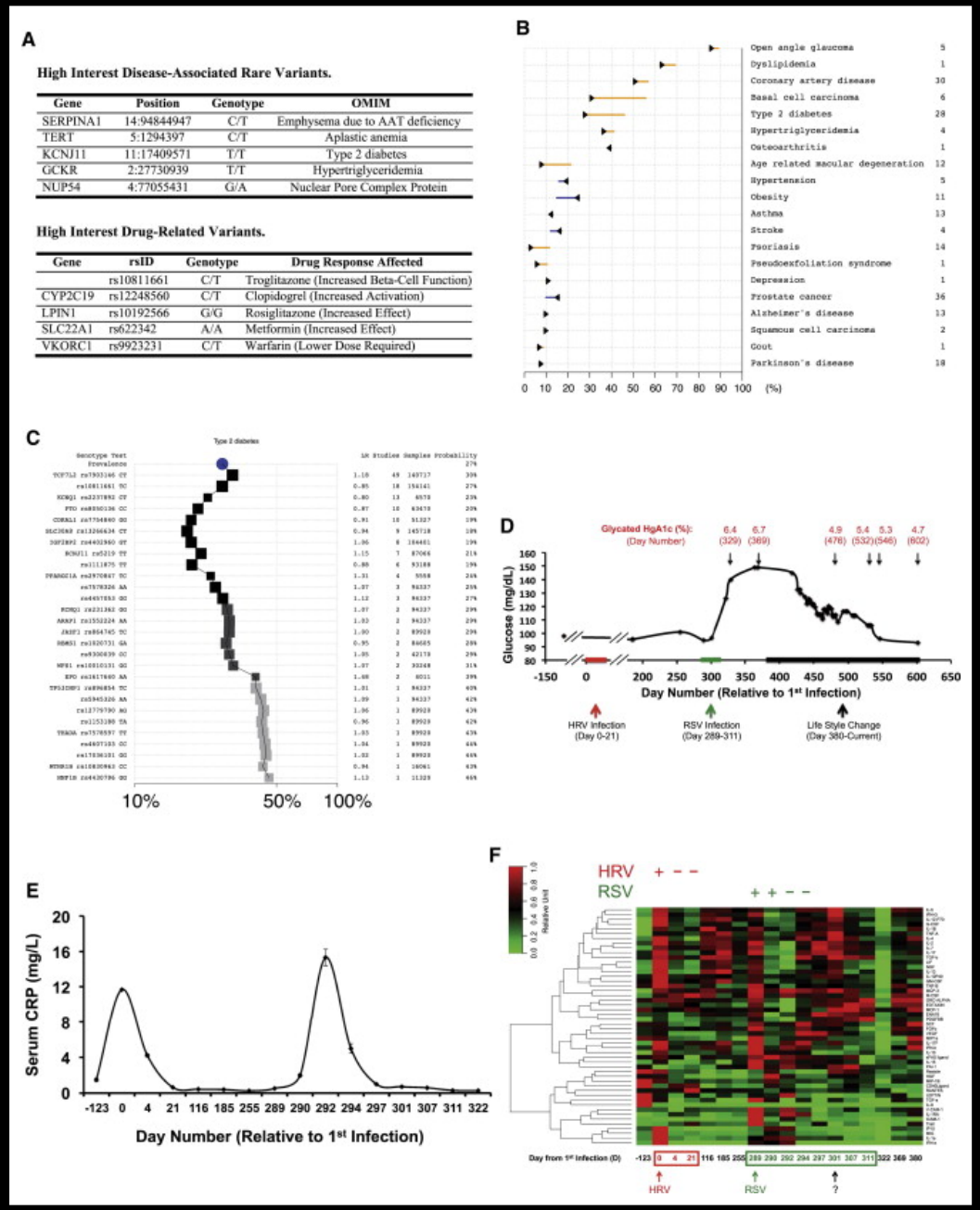
Cell PRESS
Copyright © 2012 Elsevier Inc. Terms and Conditions

Figure 1



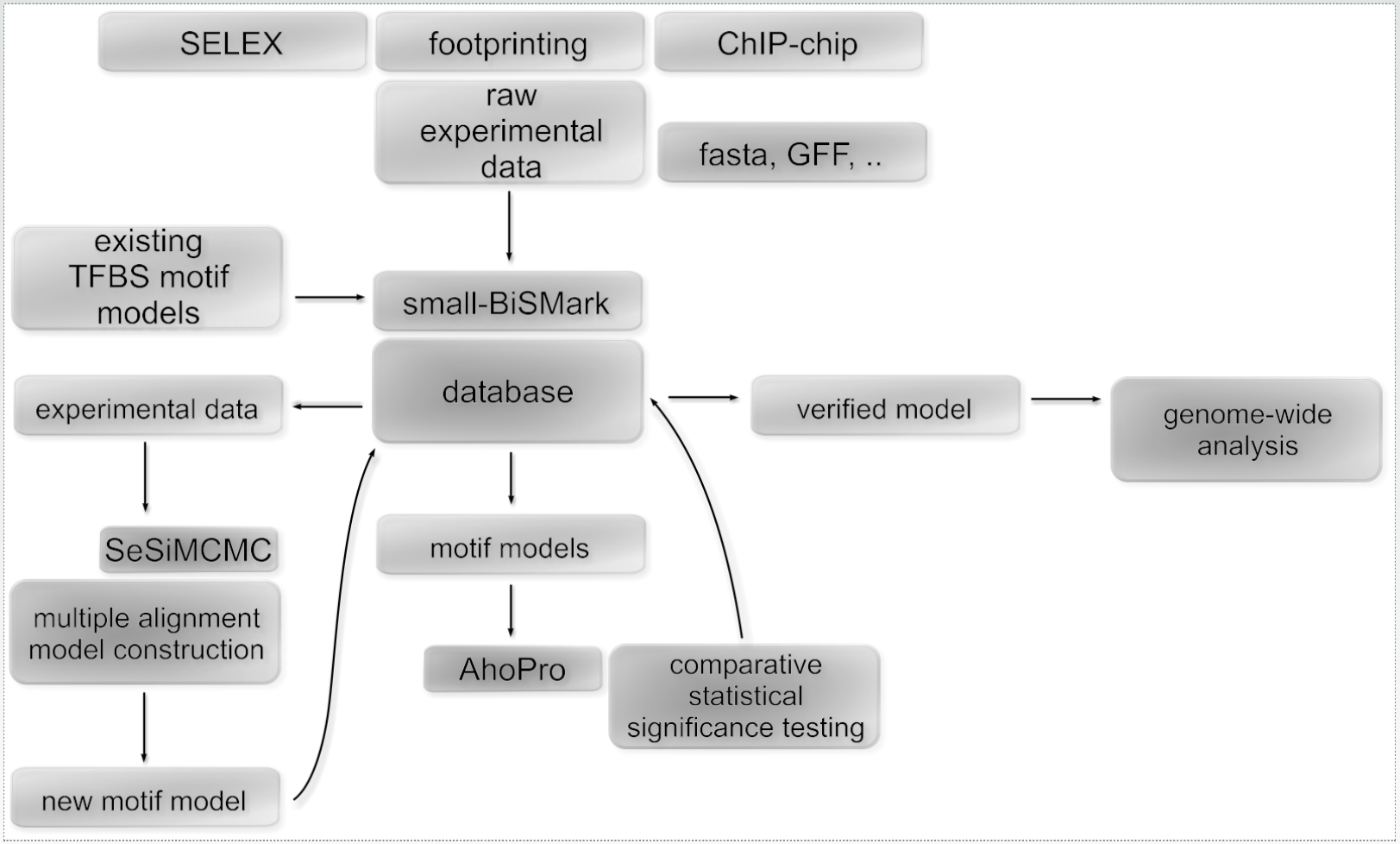
Source: Cell , Volume 148, Issue 6, Pages 1293-1307 (DOI:10.1016/j.cell.2012.02.009)
Copyright © 2012 Elsevier Inc. Terms and Conditions

Figure 2



Source: Cell , Volume 148, Issue 6, Pages 1293-1307 (DOI:10.1016/j.cell.2012.02.009)
Copyright © 2012 Elsevier Inc. Terms and Conditions

THE END


SELEX
footprinting
ChIP-chip
raw experimental data
fasta, GFF, ..
existing TFBS motif models
small-BiSMark
database
experimental data
verified model
genome-wide analysis
SeSiMCMC
motif models
multiple alignment model construction
AhoPro
comparative statistical significance testing
new motif model