Сайт связывания CTCF

В чем состоит: короткий GC богатый сигнал, до двух десятков нуклеотидов, проявляющих консервативность. Локализован в промоторах, энхансерах и изоляторах.
Кому адресован: многофункциональному белку регулятору транскрипции **CTCF**, связанным с регуляцией распространениея гетерохроматина, образованием когезиновых петель, геномным импринтингом и подавлением и стимуляцией транскрипции генов.
Предназначение: вызывает связывание белка **CTCF** с ДНК, что позволяет ему взаимодействовать с другими белками скэффолда и ремоделирования хроматина. Сила связывания зависит от метилирования ДНК и нуклеосомного окружения.
Сила сигнала: сила сигнала для сайтов связывания CTCF - это количественный показатель вероятности обнаружения сайта в определенном участке генома.
Примеры сигнала: у человека в геноме обнаружено 26000 кандидатов в сайты связывания CTCF.

1.Область контроля импринтинга H19/IGF2 (ICR) - сайт связывания CTCF, который участвует в регуляции генов H19 и IGF2, которые импринтируются и экспрессируются специфичным для родителя способом. CTCF связывается с ICR на материнской аллели и блокирует взаимодействие между промотором IGF2 и энхансером, расположенным ниже по потоку от H19, что приводит к экспрессии H19 и подавлению IGF2. [4]

2. Сайты связывания CTCF были идентифицированы в регуляторных областях гена MYC, который является протоонкогеном и часто сверхэкспрессируется при раке. CTCF может действовать как репрессор транскрипции MYC, связываясь с участком в промоторе и привлекая корепрессоры. [3]

Использованные источники: 1, 2, 3, 4

1. Tikhonova, P.; Pavlova, I.;Isaakova, E.; Tsvetkov, V.; Bogomazova,A.; Vedekhina, T.; Luzhin, A.V.; Sultanov,R.; Severov, V.; Klimina, K.; et al. DNA G-Quadruplexes Contribute to CTCF Recruitment. Int. J. Mol. Sci. 2021,22.DOI:10.3390/ijms22137090
2. CTCF binding site classes exhibit distinct evolutionary, genomic,epigenomic and transcriptomic features Kobby Essien¤*, Sebastien Vigneau¤†, Sofia Apreleva¤*, Larry N Singh*,Marisa S Bartolomei† and Sridhar Hannenhalli.DOI:10.1186/gb-2009-10-11-r131
3. Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome Tae Hoon Kim,Ziedulla K. Abdullaev,Andrew D. Smith,Michael Q. Zhang, Victor V,Lobanenkov.https://doi.org/10.1016/j.cell.2006.12.048

4. Role of CTCF Binding Sites in the Igf2/H19 Imprinting Control Region.Shih-Huey E Tang,Francisco J Silve,Walter M K Tsark.DOI:10.1128/MCB.24.11.4791-4800.2004

Нахождение последовательности представителей для сигнала в геноме и построение PWM с оценкой результаты поиска по этой PWM новых сайтов

Для анализа я выбрала сигнал Kozak в геноме человека, окрестность ATG кодона — старта транскрипции. Ссылка на позитивную выборку, на негативную (бактерия). Скрипт на colab. В качестве позитивной и тестовой выборок были выбраны старт-кодоны из белок-кодирующих транскриптов человека с окружением от -6 до +5 нуклеотидов включительно. В качестве негативной взяты все последовательности, включающие в том же положении ATG триплет из генома Echerichia coli, т.к. у прокариот последовательность Kozak встречаться не должна.

```
A    0.271199
T    0.217364
G    0.307754
C    0.203683
dtype: float64
```

	A	T	G	C
0	-0.191888	-0.063612	0.024059	0.232280
1	-0.266404	-0.030814	-0.087963	0.385958
2	-0.157137	-0.310874	-0.058342	0.447503
3	0.172074	-0.285032	0.037086	-0.043333
4	0.011456	-0.256015	-0.203024	0.407689
5	-0.220341	-0.604559	0.015912	0.544209
6	1.304902	-12.096883	-12.444613	-12.031877
7	-12.318165	1.526184	-12.444613	-12.031877
8	-12.318165	-12.096883	1.178454	-12.031877
9	-0.094543	-0.190693	0.176614	0.013387
10	-0.045832	-0.151130	-0.218292	0.409074

Рис.1.Позиционно-весовая матрица (PWM) со значением 0,1 для псевдосчётов и нуклеотидные содержания по позициям.

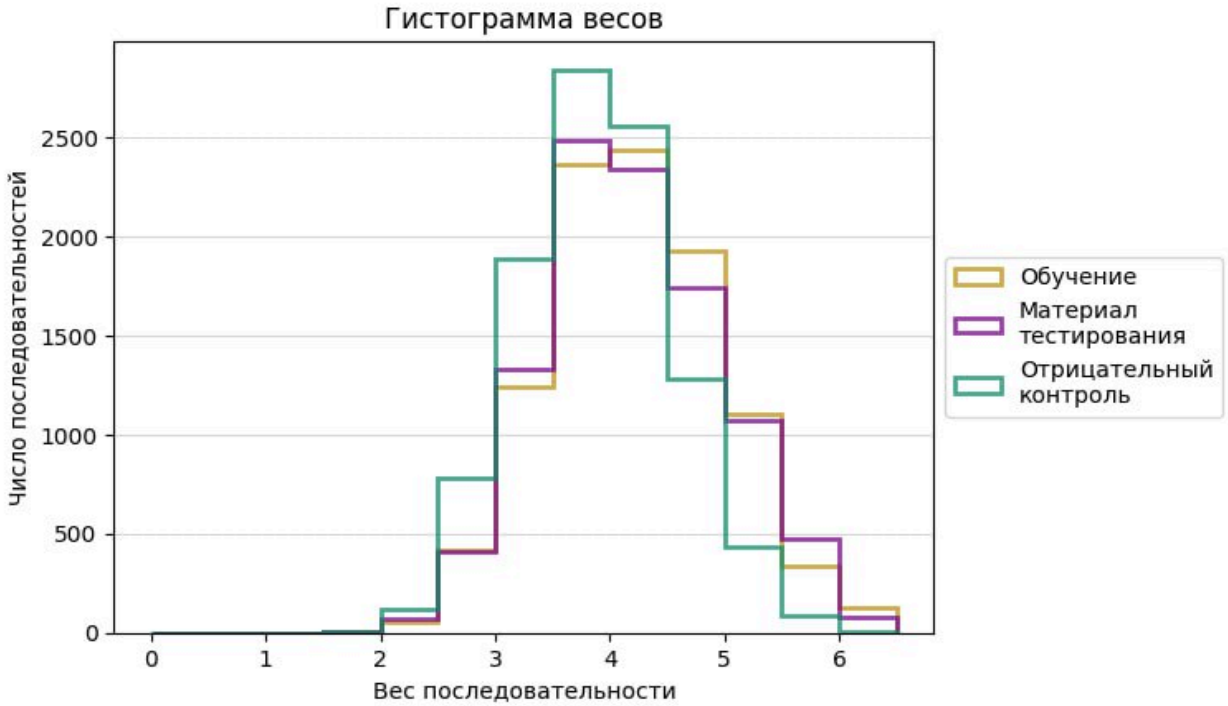


Рис.1. Диаграммы распределений весов в последовательностях. Негативная выборка с выравниванием по ATG триплету.

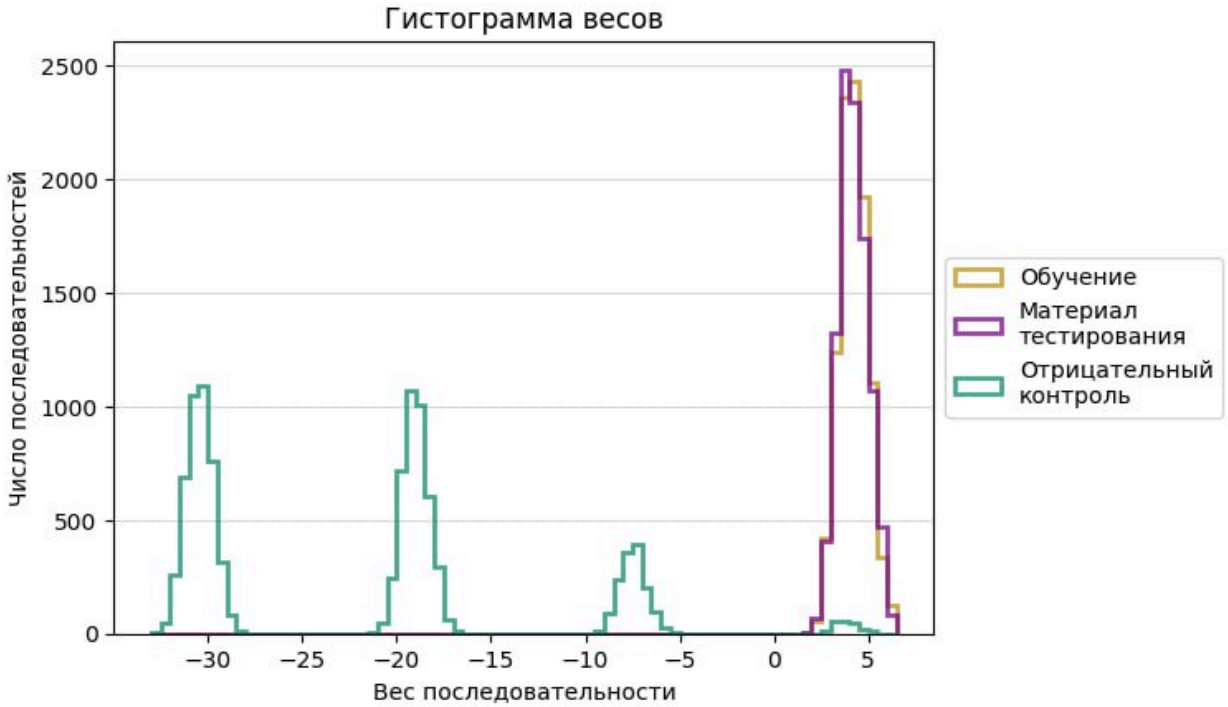


Рис.2. Диаграммы распределений весов для негативной выборки без выравнивания по ATG триплету.

Видно разбиение на 4 кластера, вероятно по присутствию нуля, одного, двух либо трех верных нуклеотидов в позициях ATG.

Исходя из медианного и среднего значений силы сигнала для негативной и тестовой выборок было выбрано пороговое значение ~3.9 и получена таблица:

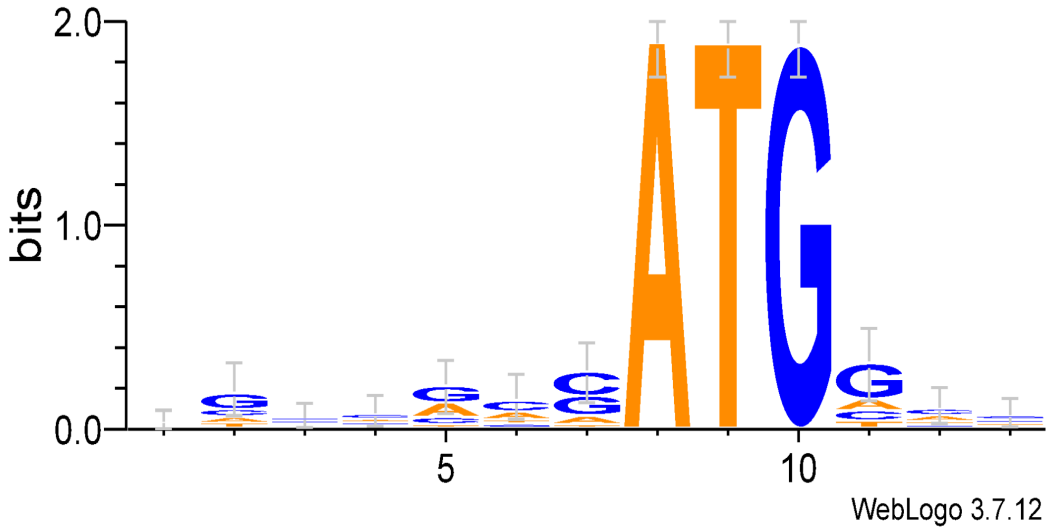
	Learn	Test	Negative
Есть сигнал	6409	6217	4924
Нет сигнала	3591	3783	5076

Таблица 1. Количество последовательностей, в которых был обнаружен сигнал.

	Learn	Test	Negative
Есть сигнал	8283	8193	135
Нет сигнала	1717	1807	9865

Таблица 2. Количество последовательностей, в которых был обнаружен сигнал с пороговым значением ~3.5 для случайной негативной выборки.

LOGO последовательности Козак человека было сгенерировано на сервере Web LOGO 3:



В результате проделанной работы была построена PWM для последовательности Козак человека. Благодаря анализу на негативной выборке удалось подтвердить консервативность (слабую) данной последовательности относительно ATG триплетов являющихся старт-кодонами человека. Можно предположить, что при выравнивании

негативной выборки по старт-кодонам бактерии, а не любым ATG триплетам, разбиение по силе сигнала окажется еще менее селективным. К сожалению, NC