АНАЛИЗ РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА ПО МАССОВОЙ ИММУНОПРЕЦИПИТАЦИИ ХРОМАТИНА С ПОМОЩЬЮ МЕТОДОВ РАСПОЗНАВАНИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ

В.Г. Левицкий, Г.В. Васильев, Д.Ю. Ощепков, Н.И. Ершов, Т.И. Меркулова

Учреждение Российской академии наук Институт цитологии и генетики Сибирского отделения РАН, Новосибирск, Россия, e-mail: levitsky@bionet.nsc.ru

Метод иммунопреципитации хроматина с последующим секвенированием всего пула выделенных фрагментов (ChIP-Seq) широко используется для получения картины полногеномного распределения сайтов связывания различных транскрипционных факторов. В настоящей работе проведен анализ структуры профилей локусов ChIP-Seq, полученных в экспериментах по иммунопреципитации хроматина печени мыши с использованием антител к транскрипционному фактору FoxA2 (Wederell *et al*., 2008), с помощью разработанных нами методов распознавания сайтов связывания белков семейства FoxA. По результатам анализа предложена следующая классификация профилей локусов: 1) унимодальные локусы (локусы, в профиле которых имеется единственный пик) длиной до 600 нт; эти локусы, вероятно, образованы одним сайтом связывания FoxA и 2) мультимодальные локусы (локусы, в профиле которых можно выделить два и более отдельно стоящих пика) длиной до 600 нт и все локусы большей длины; локусы этой группы, по-видимому, сформированы множеством отдельных сайтов.

Ключевые слова: транскрипционные факторы FoxA, сайты связывания, компьютерные методы распознавания, данные иммунопреципитации хроматина.

Введение

Среди методов, позволяющих изучать связывание транскрипционных факторов (ТФ) с ДНК, особое место занимает метод иммунопреципитации хроматина (ChIP, Farnham, 2009). Метод состоит в обработке живых клеток формальдегидом, вызывающей образование ковалентных сшивок между ДНК и близкорасположенными участками белков, а также белок-белковых сшивок. Затем хроматин дробится и с помощью иммунопреципитации со специфическими антителами выделяются районы ДНК, с которыми связываются интересующие исследователя белки (Dedon *et al*., 1991). Принципиальными преимуществами перед другими вариантами метода обладает иммунопреципитация хроматина с последующим секвенированием всего пула выделенных фрагментов (ChIP-Seq) на приборах массового параллельного секвенирования ДНК (Illumina или SOLiD). Во-первых, результатом ChIP-Seq является картина полногеномного распределения сайтов связывания транскрипционных факторов (ССТФ). Во-вторых, полученный результат свободен от предварительной селекции исходных данных, которые могут существенно искажать конечный результат, что неизбежно в случае анализа заранее выбранных последовательностей. В-третьих, результатом ChIP-Seq являются не относительные уровни сигнала, как в случае ряда других вариантов ChIP, а конкретные районы последовательностей ДНК, что дает несравненно больше возможностей для теоретического анализа результатов эксперимента. После первичной обработки результат ChIP-Seq эксперимента оказывается представленным в виде наложенной на геномную последовательность совокупности пиков, соответствующих районам концентрации от-

дельных коротких фрагментов – чтений (reads), каждый из которых является результатом единичного акта связывания. В идеальном случае пик должен иметь форму, близкую к форме кривой нормального распределения, и важной характеристикой пика является высота – число чтений, перекрывающихся на одном нуклеотиде (Robertson *et al*., 2007). Поэтому большую высоту пика принято интерпретировать либо как более высокое сродство ТФ к соответствующему сайту, либо как наличие нескольких близкорасположенных сайтов в данном районе (Jothi *et al*., 2008). Однако важно отметить, что обычно значительная часть пиков ChIP-Seq имеет форму, существенно отличающуюся от теоретически ожидаемой. Например, пик может иметь несколько вершин вместо одной, иметь значительную длину (несколько т.п.н.) и т. д., что допускает несколько различных объяснений. Во-первых, это может быть результатом статистических флуктуаций, вызванных недостаточной глубиной чтения. Во-вторых, к аналогичному эффекту может приводить кластер близко расположенных ССТФ. В-третьих, на особенности позиционирования ТФ на ДНК существенное влияние могут оказывать структура хроматина и взаимодей-ствие данного ТФ с ТФ-партнерами и другими компонентами хроматина. Одним из подходов, позволяющих приблизиться к решению этих вопросов, являются надежные биоинформатические методы распознавания ССТФ в последовательностях ДНК подпиковых областей.

Целью данной работы были исследование формы пиков ChIP-Seq, полученных в экспериментах по иммунопреципитации хроматина печени мыши с использованием антител к ТФ FoxA2 (Wederell *et al*., 2008), и определение плотности и локализации потенциальных FoxA сайтов в подпиковых районах локусов разной формы профиля.

Материалы и методы

Разработка критериев сортировки профилей локусов ChIP-Seq эксперимента

Данные ChIP-Seq эксперимента по выявлению мишеней FoxA2 взяты из работы (Wederell *et al*., 2008, http://www.bcgsc.ca/data/ChIP-Seq). Они представлены в виде: 1) профиля ChIP-Seq эксперимента, который содержит целые значения ≥ 1 и показывает, сколько раз происходило наложение секвенированных участков ДНК на соответствующий район геномной последовательности; 2) таблицы хромосомных локализаций 11475 районов геномной ДНК мыши (локусов), для которых максимум профиля ChIP-Seq был равен или превышал значение 10. Авторы полагают, что именно это значение позволяет надежно выявить районы, потенциально содержащие мишени ТФ FoxA. Поэтому далее аннотированным пиком мы будем считать область локуса с высотой профиля ChIP-Seq не менее 10. Максимальная высота пика по всем локусам составляет 251.

Профиль считается унимодальным, если в нем есть одна неразрывная область максимальных значений (единственный пик), в мультимодальном профиле можно выделить два и более отдельно стоящих пика. Для определения, является ли профиль локуса уни- или мультимодальным, нами применен следующий подход. Пусть профиль длины L нт представляется значениями $\{H_n\}$, $1 \leq n \leq L$. Выберем некоторое значение ширины Δ нт для расчетов локальных максимумов профиля $\{M_n\}$. Если $\Delta < L$, то при условии $\Delta/2 > n > L - \Delta/2 + 1$ $M_n = \underset{i = n - \Delta/2 + 1}{\overset{n + \Delta/2}{Max}}(H_i)$, а в случаях $n < \Delta/2$ или $n > L - \Delta/2 + 1$ соответственно $M_n = M_{\Delta/2}$ или $M_n = M_{L - \Delta/2 + 1}$. Если $\Delta \geq L$, то для любой позиции профиля: $M_n = \underset{i = 1}{\overset{L}{Max}}(H_i)$. Расчет профиля локальных максимумов $\{M_n\}$ и выбор порога отсечения T $(0 < T < 1)$ позволяют разделить все L позиций профиля на высокие «холмы» и низкие «ямы». Если значение профиля H_n удовлетворяет условию $H_n \geq T \times M_n$, то эта n-я позиция определялась нами как «холм», в противном случае – как «яма». При заданных параметрах Δ и T унимодальным считается локус, имеющий один непрерывный холм, в противном случае локус считается мультимодальным. В анализе мы использовали одно значение интервала $\Delta = 500$ нт и два значения порога отсечения $\{T_1, T_2\} = \{0{,}5, 0{,}8\}$. Выборки уни- и мультимодальных локусов составлялись на основе критерия сохранения уни- или мультимодальности локуса при двух порогах отсечения T_1 и T_2.

Выборки уни- и мультимодальных локусов, использованные в анализе

По исходной выборке 11475 локусов (Wederell *et al*., 2008) с применением описанного выше подхода составлены выборки всех унимодальных и мультимодальных локусов объемов 4941 и 2909 соответственно. Наряду с этими двумя классами в исходной выборке выделяются два смешанных класса: 1) 1977 локусов, которые при пороге отсечения 0,5 классифицировались как унимодальные, а при пороге 0,8 как мультимодальные; 2) 1648 локусов, каждый из которых при этих же порогах являлся мульти- и унимодальным соответственно. Далее смешанные классы мы не рассматривали.

Для анализа обогащения потенциальных ССТФ в локусах (а) вблизи пиков; (б) с различной максимальной высотой пика и (в) сравнения уни- и мультимодальных локусов нами были составлены следующие выборки: районы «холмов» уни- и мультимодальных локусов, максимальная высота пика для которых не превышала 10, 14 и 20. Для унимодальных локусов, по их определению, один «холм» соответствует одному локусу, в случае мультимодальных из каждого локуса в выборку включался единственный «холм» с самой большой высотой пика, в случае равенства высот двух «холмов» выбирался один, более протяженный. Всего составлено 12 выборок: по 3 уни- и мультимодальных типа для двух порогов отсечения 0,8 и 0,5. В табл. 1 дано описание составленных выборок уни- и мультимодальных локусов.

Выборки всех уни- и мультимодальных локусов с высотой пика ≥ 10 были разбиты соответственно на 6 и 5 классов согласно распределению длин последовательностей (табл. 2). В каждую из выборок входили только последовательности, длины которых удовлетворяют определенным условиям (1-я колонка табл. 2).

Разработка метода распознавания FoxA сайтов

Для составления обучающих выборок ССТФ подсемейства FoxA была использована выборка последовательностей 81 экспериментально подтвержденного сайта. Часть из них, 37 последовательностей, была взята из базы данных SAMPLES (http://srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-page+LibInfo+-newId+-lib+SAMPLES) (Kolchanov *et al*., 2002), а остальные 44 последовательности были собраны по литературным данным.

Для выявления наиболее представительных вырожденных олигонуклеотидных мотивов – коротких слов фиксированной длины, записанных в 15-буквенном IUPAC коде, в обучающей выборке использовался метод генетического алгоритма (ГА) (Levitsky, 2010). Мы произвели поиск набора мотивов длины не менее 12 нт, при этом допускалось описание одного сайта сразу несколькими мотивами. Затем на основе выявленных мотивов были составлены две выравненных обучающих выборки ССТФ FoxA. Выборки были использованы для построения методов распознавания ССТФ FoxA на основе подхода

Таблица 1

Выборки районов «холмов» уни- и мультимодальных локусов, полученные при двух порогах отсечения *T* профилей ChIP-Seq

Тип выборки	Высота пика	Объем выборки	Медиана по длинам последовательностей в выборке (нт) при пороге отсечения *T*	
			T = 0,8	*T* = 0,5
Унимодальные	≥ 10	4941	137	242
	≥ 14	2329	139	243
	≥ 20	1129	144	245
Мультимодальные	≥ 10	2909	130	245
	≥ 14	1312	133	243
	≥ 20	530	140	242

Таблица 2

Выборки уни- и мультимодальных локусов, соответствующие сходным диапазонам длин локусов L (для всех выборок высота пика ≥ 10)

Диапазон длины, *L* (нт)	Уни-модальные	Мульти-модальные
$L \leq 400$	1291	
$400 < L \leq 600$	1933	183
$600 < L \leq 800$	1239	395
$800 < L \leq 1000$	397	744
$1000 < L \leq 1200$	75	689
$L > 1200$	6	898
∑	4941	2909

оптимизированных весовых матриц (PWM, Position Weight Matrix) (Levitsky *et al*., 2007). Ранее установлено, что повышению точности распознавания способствуют (1) привлечение фланкирующих последовательностей и (2) рассмотрение динуклеотидной статистики (Zhang, Marr, 1993; Gershenzon *et al*., 2005; Levitsky *et al*., 2007). Поэтому в каждой из выборок длина последовательностей, включающих мотив в центральном положении, составляла 32 нт, а для расчета весов матриц мы использовали динуклеотидные частоты. Для анализа результатов ChIP-Seq эксперимента (Wederell *et al*., 2008) с помощью 1-й и 2-й матриц были выбраны пороги 0,68 и 0,685 соответственно. Эти пороги установлены методом оценки точности распознавания со скользящим контролем (jack-knife), они соответствуют ошибке недопредсказания 50 % и ошибкам перепредсказания 3,8E–04 и 5,1E–04 (Левицкий и соавт., в печати).

Результаты и обсуждение

Построение методов распознавания ССТФ FoxA на основе поиска мотивов

Проведенный ранее предварительный анализ последовательностей ССТФ FoxA выявил большую их вариабельность. Так, например, в выборке из 81 экспериментально подтвержденного ССТФ FoxA с помощью выявленного ранее консенсуса сайта FoxA VAWTRTTKRTY (Welsheimer, Newbold, 1996) удалось обнаружить лишь 18 последовательностей с не более чем одним несовпадением, что составило около 22 % от всей выборки. В связи с этим составление стандартными методами выравнивания качественной выборки, включающей все сайты (81), было невозможным. Очевидно, требовалось выделение подвыборок сходных сайтов из исходной выборки. Эта задача была решена путем поиска вырожденных олигонуклеотидных мотивов. В качестве наиболее представленных мотивов было найдено два: TRTTTRYH и YRTTKDYYYD (R=A/G, Y=T/C, H=A/T/C, K=T/G, D=A/T/C). Мотив TRTTTRYH встретился в 53 последовательностях выборки (более 65 % от всего объема). Мотив YRTTKDYYYD встретился в другом (частично пересекающемся с первым) наборе из 48 последовательностей (более 59 %). Мотивы TRTTTRYH и YRTTKDYYYD вместе или отдельно содержатся в 64 последовательностях выборки из 81 ССТФ FoxA (79 %) и, следовательно, существенно лучше для описания выборки, чем консенсус VAWTRTTKRTY (Welsheimer, Newbold, 1996), описывающий лишь 22 % выборки. При этом оценка частот встреч по случайным причинам мотива VAWTRTTKRTY с не более чем одним несовпадением составляет 1/1021 нт, а для мотивов TRTTTRYH и YRTTKDYYYD эти оценки составляют 1/1183 и 1/813 нт соответственно. Таким образом, при значительно лучшем описании выборки парой выявленных нами мотивов по сравнению с принятым консенсусом вероятность наблюдения каждого из рассматриваемых мотивов по случайным причинам оказывается сопоставимой. На основании найденных мотивов было сформировано две выравненных подвыборки ССТФ FoxA, которые легли в основу построения методов распознавания этих факторов и были обозначены как «метод D» и «метод E», использованные для анализа данных ChIP-Seq эксперимента.

Общий анализ выборок ChIP-Seq локусов

Для изучения обогащения потенциальными сайтами FoxA как всего набора локусов ChIP-Seq, так и последовательностей, классифицированных нами как «холмы» в уни- и мультимодальных локусах при двух порогах отсечения

T = 0,5 и 0,8, было проведено распознавание сайтов FoxA с помощью метода D (табл. 3). Для оценки обогащения мы сравнивали плотность потенциальных сайтов в локусах с плотностью потенциальных FoxA в геноме мыши и в случайных последовательностях, полученных путем случайного многократного перемешивания букв в последовательностях локусов. Аналогичные результаты были получены с использованием метода E.

Как видно из табл. 3, в целом геном обогащен потенциальными сайтами FoxA по сравнению со случайными последовательностями. Как ожидалось, в холмах как уни-, так и мультимодальных локусов плотность потенциальных сайтов существенно выше, чем в полноразрных локусах. В свою очередь, плотность потенциальных сайтов в полноразмерных локусах заметно выше, чем в полном геноме и случайных последовательностях. Также важно отметить, что чем жестче порог отсечения (T = 0,8 по сравнению с T = 0,5), тем плотность сайтов выше, что говорит в целом о том, что потенциальные сайты предпочтительно располагаются ближе к максимуму пика.

Доля локусов, содержащих потенциальные FoxA сайты, и анализ отклонений позиций потенциальных сайтов от аннотированных пиков

Учитывая, что в ходе эксперимента ChIP-Seq для проведения массового секвенирования геномная ДНК разрезается на фрагменты средней длины 200 нт (Wederell *et al*., 2008), то возможно смещение пиков относительно сайтов FoxA на расстояние до 200 нт. Именно поэтому мы можем принять отклонение 200 нт как допустимое смещение потенциальных сайтов от пиков профиля ChIP-Seq. Напомним, что аннотированным

Таблица 3

Оценка обогащения потенциальными сайтами FoxA (метод D) последовательностей ChIP-Seq по сравнению с ожиданием по полному геному мыши и по случайным последовательностям

Выборки Background			Частота встреч	Отношение частоты встреч по отношению:	
				к случайным последовательностям	к геномным последовательностям
Случайные			1/3512	1	0,34
Геном (все хромосомы мыши)			1/1178	2,98	1
Выборки ChIP-Seq		Высота пика			
Вся длина	Все локусы	≥ 10	1/765	4,59	1,54
	Все унимодальные		1/711	4,94	1,66
	Все мультимодальные		1/834	4,21	1,41
Холмы, T = 0,8	Унимодальные	≥ 10	1/302	11,64	3,90
		≥ 14	1/247	14,20	4,76
		≥ 20	1/226	15,51	5,20
	Мультимодальные	≥ 10	1/305	11,51	3,86
		≥ 14	1/288	12,19	4,09
		≥ 20	1/275	12,78	4,29
Холмы, T = 0,5	Унимодальные	≥ 10	1/469	7,48	2,51
		≥ 14	1/406	8,64	2,90
		≥ 20	1/369	9,50	3,19
	Мультимодальные	≥ 10	1/499	7,03	2,36
		≥ 14	1/480	7,32	2,45
		≥ 20	1/466	7,54	2,53

пиком мы считаем область локуса с высотой профиля ChIP-Seq не менее 10. Таким образом, нами были приняты следующие критерии для подтверждения достоверности распознавания предсказанных методом PWM потенциальных сайтов на основе данных профилей ChIP-Seq: 1) каждое предсказание должно иметь значение профиля PWM больше заданного порога; 2) отклонение позиции потенциального сайта от позиции профиля со значением большим либо равным 10 должно быть не более 200 нт.

Проанализированы выборки полноразмерных локусов, для которых максимальная высота пика не превосходит заданное значение $H_m \geq H$, ($10 \leq H \leq 50$). Для каждой выборки определены: 1) отношение числа локусов, в которых есть хотя бы один потенциальный сайт, к числу всех локусов; 2) отношение числа локусов, в которых есть хотя бы один потенциальный сайт с отклонением не более 200 нт от ближайшей позиции локуса с высотой ≥ 10, к числу всех локусов. Результаты расчетов представлены на рис. 1 и 2. Данные расчеты проводились для методов D, E, а также для метода, построенного на основе их объединения, т. е. в этом случае учитывались сайты, распознанные любой из двух матриц.

Как видно из рис. 1, 2, методы D и E описывают меньшее количество сайтов в локусах с большой высотой пика по сравнению с локусами со средней высотой пика. Максимальную долю локусов методы D и E описывают в диапазоне высот пиков от 25 до 30. Из сравнения рис. 1 и 2 можно показать, что в 7 % локусов с максимальной высотой пика не менее 10 потенциальный сайт отклоняется от аннотированного пика более чем на 200 нт. При возрастании высоты пика до 25–30 это значение падает до 4 %, снижаясь до 2 % на больших высотах пика (около 50).

Анализ отклонений позиций потенциальных сайтов от аннотированных пиков для уни- и мультимодальных локусов

Рассмотрим две выборки локусов, уни- и мультимодальных, и проведем расчет отклонений потенциальных сайтов от аннотированных пиков. Затем по каждой из этих выборок мы рассчитаем долю потенциальных сайтов, для которых отклонение от аннотированного пика будет меньше заданного значения (10, 20 нт и т. д.). Результаты расчетов для метода D представлены

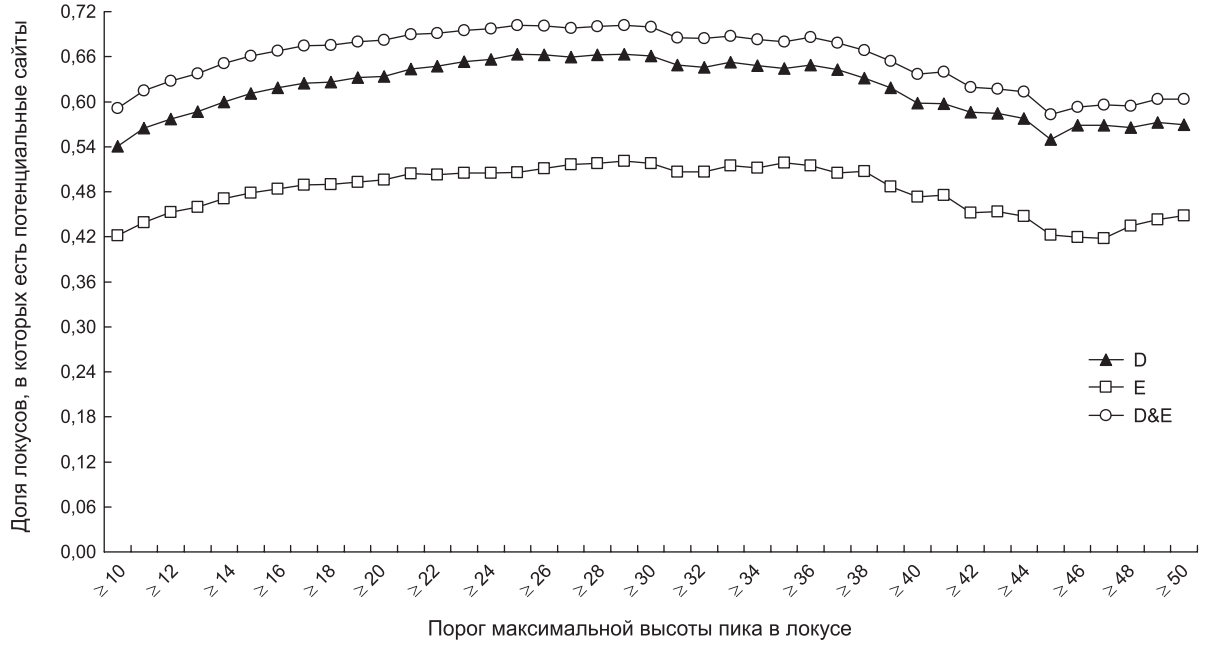

Рис. 1. Анализ выборок локусов, характеризуемых разной максимальной высотой пиков.

Каждая точка графика соответствует расчетам для выборки, в которую входили только локусы с максимальной высотой пика не менее заданного значения (ось X). По оси Y отложена доля локусов, имеющих потенциальные сайты, по отношению к числу всех локусов для заданной выборки.

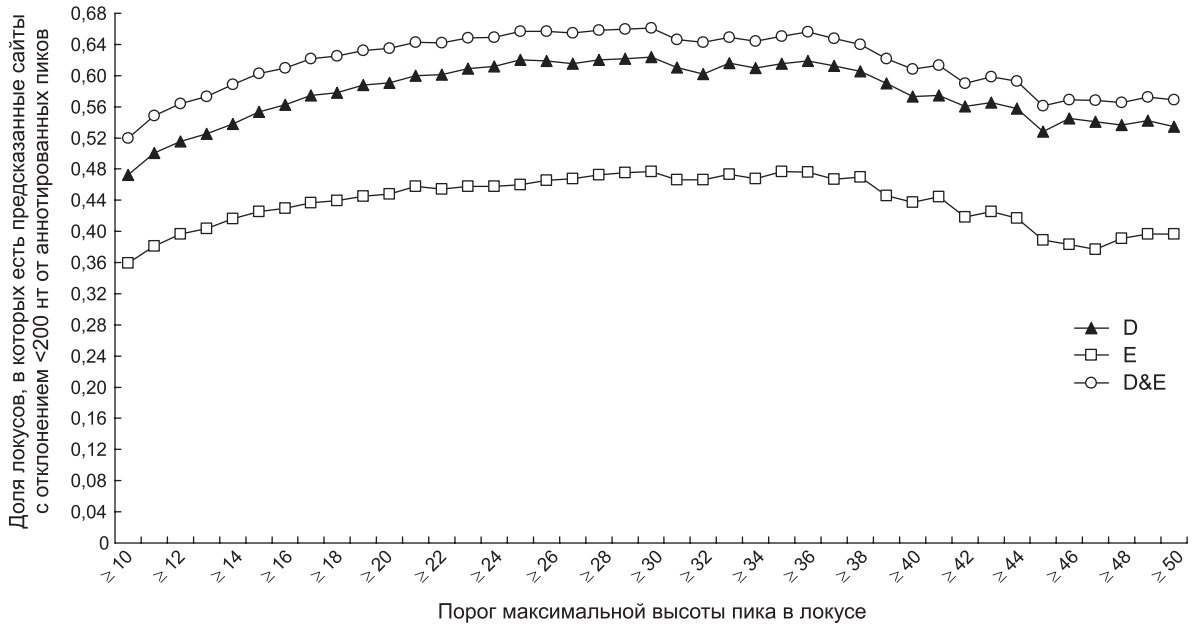

Рис. 2. Анализ выборок локусов, характеризуемых разной максимальной высотой пиков.

Каждая точка графика соответствует расчетам для выборки, в которую входили только локусы с максимальной высотой пика не менее заданного значения (ось X). Ось Y – отношение числа локусов, имеющих потенциальные сайты с отклонением менее 200 нт от пика с высотой не менее 10, к числу всех локусов с заданной максимальной высотой.

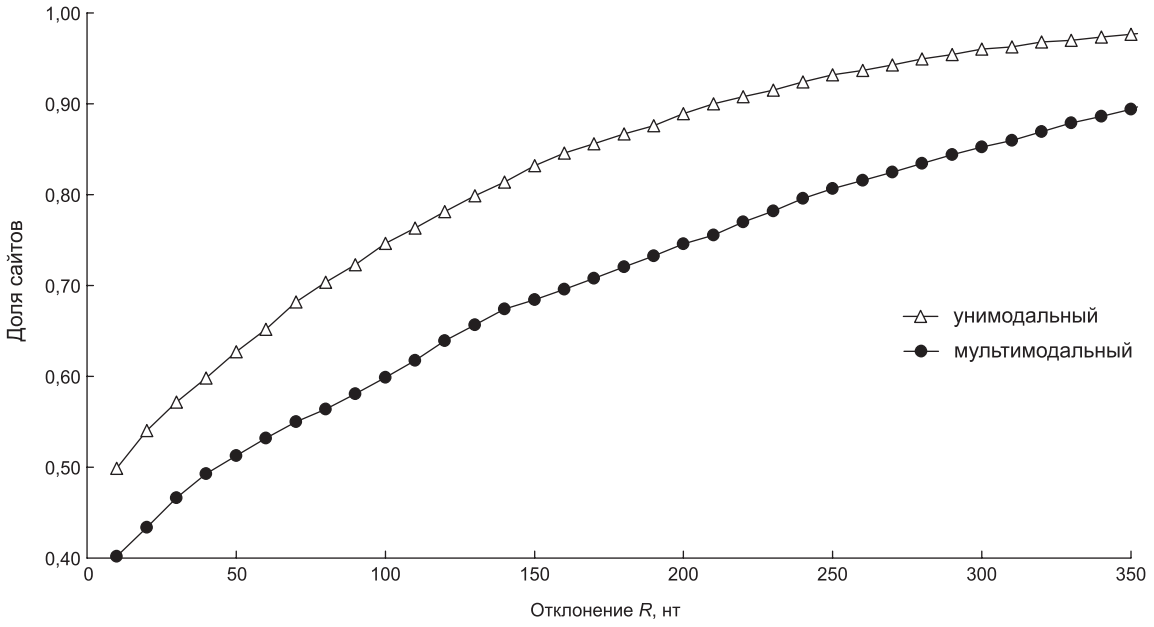

Рис. 3. Доли потенциальных сайтов в унимодальных и мультимодальных локусах с отклонением не более заданного от аннотированных пиков.

Сайты предсказаны с помощью матрицы D.

на рис. 3. Так, видно, что 74,6 % всех потенциальных сайтов в унимодальных локусах имеют отклонение либо меньшее, либо равное 100 нт, тогда как для мультимодальных локусов соответствующая доля составляет 59,9 % (рис. 3).

Для сравнения степени обогащения потенциальными сайтами районов вблизи пиков в унимодальных и мультимодальных локусах или, другими словами, для выявления достоверности связи между нахождением потенциального сайта не далее чем на расстоянии *R* нт от аннотированного пика и типом локуса нами был применен следующий подход. Для каждого значения максимального отклонения составлялась таблица сопряженности (табл. 4). На основании этой таблицы рассчитывался критерий χ^2, который позволяет оценить достоверность гипотезы о том, что в унимодальных локусах чаще, чем в мультимодальных, потенциальные сайты расположены не далее чем на расстоянии *R* нт от аннотированного пика.

Оценки значимости критерия χ^2, рассчитанные для значений отклонений от 0 до 350 нт между аннотированными пиками и сайтами, предсказанными методами D и E, приведены на рис. 4. Заметно, что для обоих методов существует обширный диапазон значений максимального отклонения, где критерий значим.

Таблица 4

Таблица сопряженности 2 × 2 для проверки достоверности связи между нахождением потенциального сайта вблизи пика и типом локуса

Тип локуса	Число сайтов		Число всех сайтов в соответствующем типе локуса
	с отклонением ≤ R от аннотированного пика	с отклонением > R от аннотированного пика	
Унимодальные	X_1	X_3	$X_1 + X_3$
Мультимодальные	X_2	X_4	$X_2 + X_4$
Всего	$X_1 + X_2$	$X_3 + X_4$	

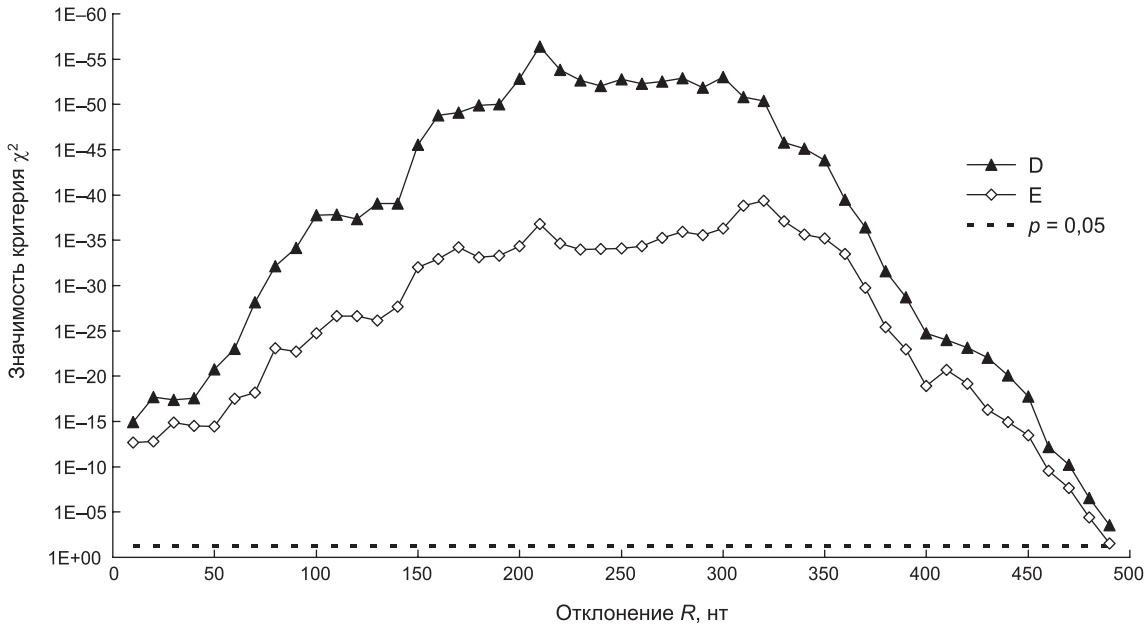

Рис. 4. Значимость по критерию χ^2 гипотезы о том, что в унимодальных локусах чаще, чем в мультимодальных, потенциальные сайты расположены не далее чем на расстоянии *R* нт от аннотированного пика.

Ось X – уровень значимости согласно критерию χ^2 для таблиц сопряженности 2 × 2. Ось Y – верхний порог отклонения *R* (нт). Сайты предсказаны с помощью методов D и E. Пунктиром указан уровень, соответствующий критической значимости $p = 0{,}05$.

Практически для любых расстояний районы вблизи пиков в унимодальных локусах обогащены потенциальными сайтами по сравнению с районами мультимодальных локусов. Наиболее значимым критерий оказывается для диапазона отклонений от 200 до 300 нт.

Из рис. 4 видно, что пик значимости для матриц D и E достигается для отклонения, равного 210 нт ($p < 5E{-}57$ и $p < 2E{-}37$ соответственно). Представим в виде гистограммы значения в ячейках матрицы сопряженности, рассчитанной так же, как было описано выше по отклонениям не более 210 нт для метода D (рис. 5).

Из рис. 5 хорошо видно, насколько чаще потенциальные сайты встречаются вблизи (не далее 210 нт) унимодальных пиков по сравнению с мультимодальными. Таким образом, можно с уверенностью утверждать, что районы вблизи аннотированных пиков в унимодальных локусах обогащены потенциальными сайтами по сравнению с районами вблизи пиков в мультимодальных локусах. Однако при этом исследовании нами до сих пор никак не учитывалась длина локуса. В то же время этот показатель весьма важен: в случае присутствия единственного сайта по способу построения профилей локус не должен значительно превышать длину 400 нт. Исходя из этих соображений анализ более протяженных локусов представляет существенный интерес. Поэтому следующим шагом мы провели анализ отклонений позиций потенциальных сайтов от аннотированных пиков, принимая во внимание длины локусов.

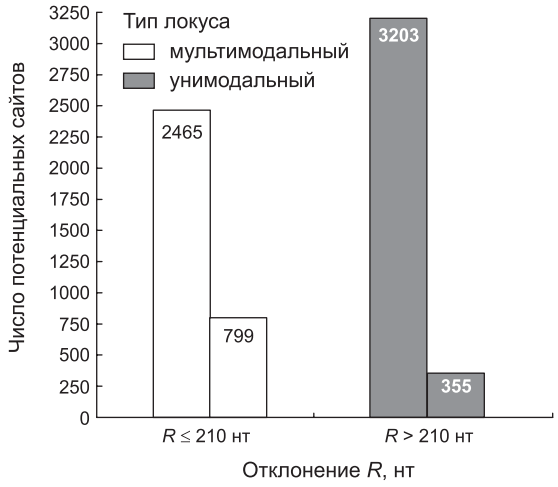

Рис. 5. Количество потенциальных сайтов, классифицированных согласно признакам: отклонение сайта от пика ≤ 210 нт или > 210 нт; тип локуса – уни- или мультимодальный.

Анализ отклонений позиций потенциальных сайтов от аннотированных пиков для уни- и мультимодальных локусов сходной длины

Был проведен расчет отклонений потенциальных сайтов от аннотированных пиков в выборках уни- и мультимодальных локусов, классифицированных нами по их длинам (табл. 2). Затем для каждой из выборок локусов мы рассчитали долю потенциальных сайтов, для которых отклонение от аннотированного пика было меньше заданного значения (10, 20 нт, и т. д.). На рис. 6 представлены результаты расчетов для сайтов, предсказанных методом D в выборках унимодальных локусов длиной до 400 нт, а также для уни- и мультимодальных локусов с длиной от 400 до 600 нт, от 600 до 800 нт и свыше 800 нт. Так, видно, что в выборках унимодальных локусов длиной до 400 нт отклонение 100 % сайтов, выявленных в этих локусах, не превышает 180 п.н. Этот результат ожидаем из чисто теоретических соображений: поскольку унимодальные пики имеют форму, близкую к симметричной, отклонение не должно превышать 400/2 = 200 нт (рис. 6). При сравнении унимодальных и мультимодальных локусов длиной от 400 до 600 нт можно отметить, что районы вблизи аннотированных пиков в унимодальных локусах этой длины обогащены потенциальными сайтами по сравнению с районами вблизи пиков в мультимодальных локусах, и этот эффект заметен в случае небольших отклонений (до 110 нт). Однако при сравнении унимодальных и мультимодальных локусов длиной более 800 нт мы не наблюдаем подобного значимого обогащения. Дальнейшие расчеты (рис. 7) показали, что значимое обогащение наблюдается только при отклонениях менее 50 нт и более 150 нт. Таким образом, можно с уверенностью утверждать, что выявленная нами связь между нахождением потенциального сайта на расстоянии не далее чем 200 нт от аннотированного пика и типом локуса в значительной степени зависит от длины самого локуса.

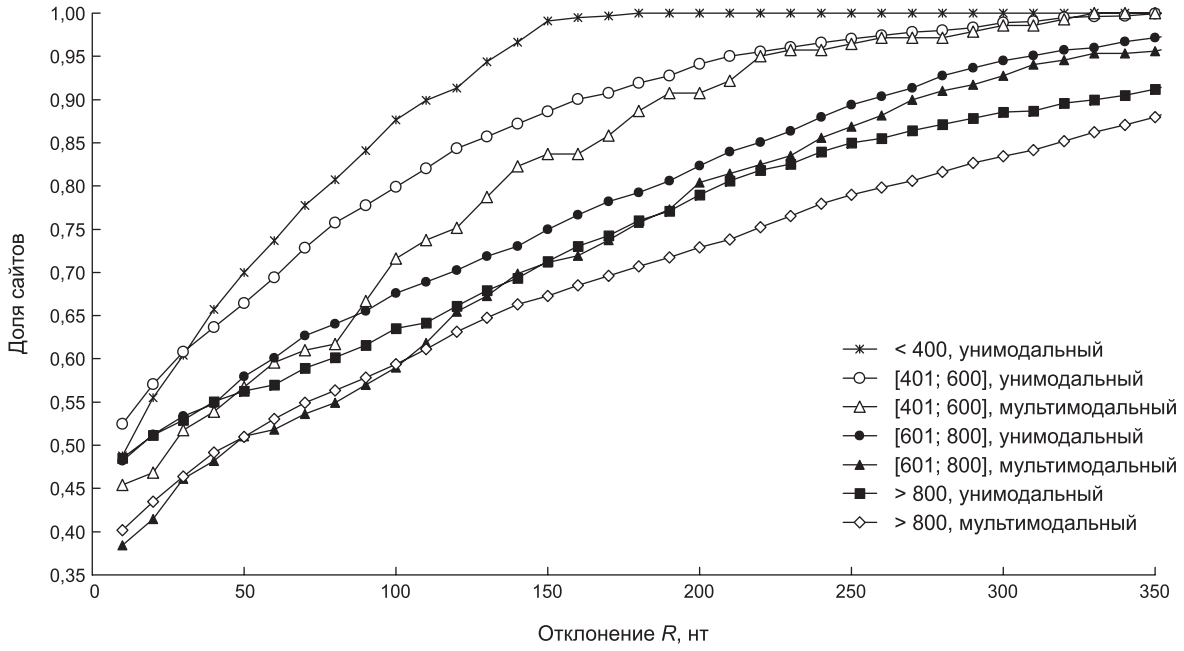

Рис. 6. Доли потенциальных сайтов в унимодальных и мультимодальных локусах различных диапазонов длин с отклонением не более заданного от аннотированных пиков.

Сайты предсказаны с помощью матрицы D.

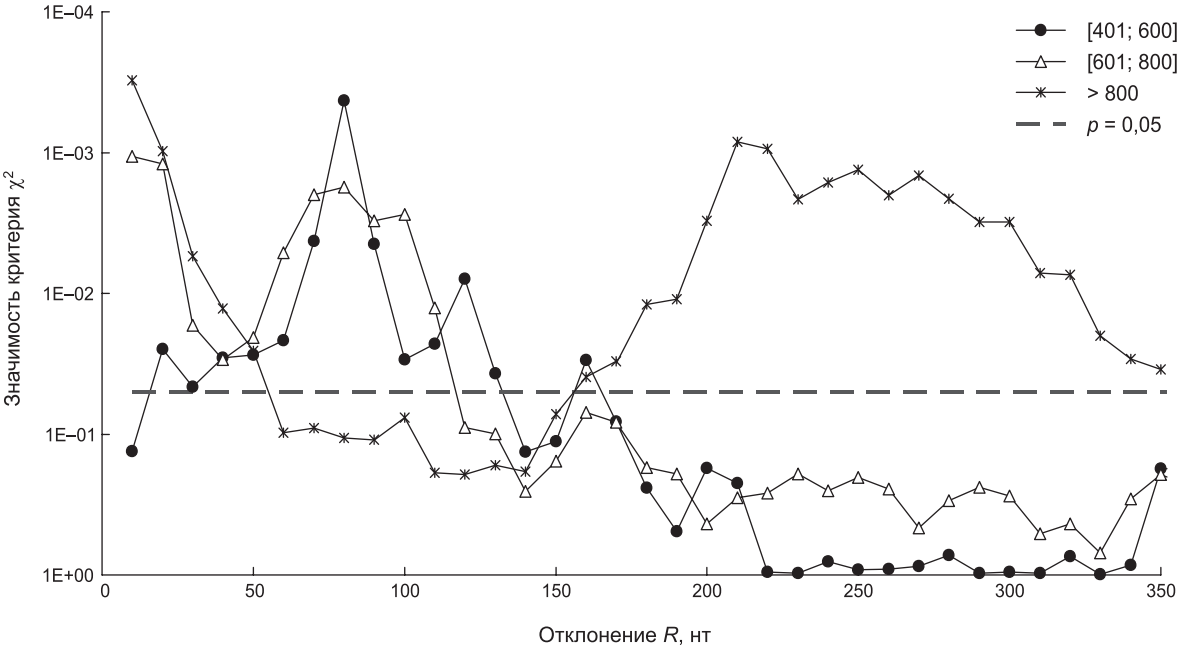

Рис. 7. Значимость по критерию χ^2 гипотезы о том, что в унимодальных локусах определенной длины чаще, чем в соответствующих мультимодальных, потенциальные сайты расположены с меньшим отклонением от аннотированных пиков.

Ось X – уровень значимости согласно критерию χ^2 для таблиц сопряженности 2 × 2. Ось Y – верхний порог отклонения R (нт). Сайты предсказаны с помощью метода D. Пунктиром указан уровень, соответствующий критической значимости $p = 0{,}05$.

Для проверки значимости выявляемого обогащения мы сравнивали между собой попарно выборки уни- и мультимодальных локусов сходной длины, описанные в табл. 2. Значимость оценивалась с помощью критерия χ^2 согласно описанной в предыдущем разделе методике с составлением таблиц сопряженности, аналогичных табл. 4. На основании этой таблицы рассчитывались значения критерия χ^2, которые позволяют оценить достоверность гипотезы о том, что в унимодальных локусах чаще, чем в мультимодальных, потенциальные сайты расположены с отклонением не более *R* нт от аннотированного пика. Оценки значимости критерия χ^2, рассчитанные для значений отклонений от 0 до 350 нт между аннотированными пиками и сайтами, предсказанными методом D для трех пар выборок уни- и мультимодальных локусов разной длины, приведены на рис. 7.

Из рис. 7, в частности, видно, что для отклонений от 10 до 130 нт районы вблизи пиков в унимодальных локусах длиной от 400 до 600 нт значимо обогащены потенциальными сайтами по сравнению с районами мультимодальных локусов того же диапазона длин. Этот результат подтверждает наблюдение, которое мы сделали по результатам анализа распределений, представленных на рис. 6.

Для локусов длиной от 600 до 800 нт феномен обогащения все также присутствует, но он незначительно ослабевает. Так, согласно представленным на рис. 7 данным, значимое обогащение наблюдается в диапазонах отклонений менее 110 нт. Для локусов длиной более 800 нт значимое обогащение районов вблизи пиков в унимодальных локусах по сравнению с мультимодальными обнаружено только при значениях отклонения менее 50 нт и более 150 нт (рис. 7), при остальных отклонениях значимости не выявлено. Таким образом, мы наблюдаем заметное ослабление феномена обогащения, с ростом длин локусов свыше 800 нт он ослабевает или вообще не наблюдается. Анализ результатов, полученных с помощью распознавания потенциальных сайтов методом E, приводит к аналогичным выводам.

Гистограммы количества потенциальных сайтов, выявленных в выборках локусов различных диапазонов длин и классифицированных согласно признакам отклонения сайта от пика ≤ 80 нт или > 80 нт и типа локуса, представлены на рис. 8, а, б. Эти гистограммы наглядно демонстрируют наличие и исчезновение феномена соответственно.

В целом проведенная дополнительная классификация уни- и мультимодальных локусов

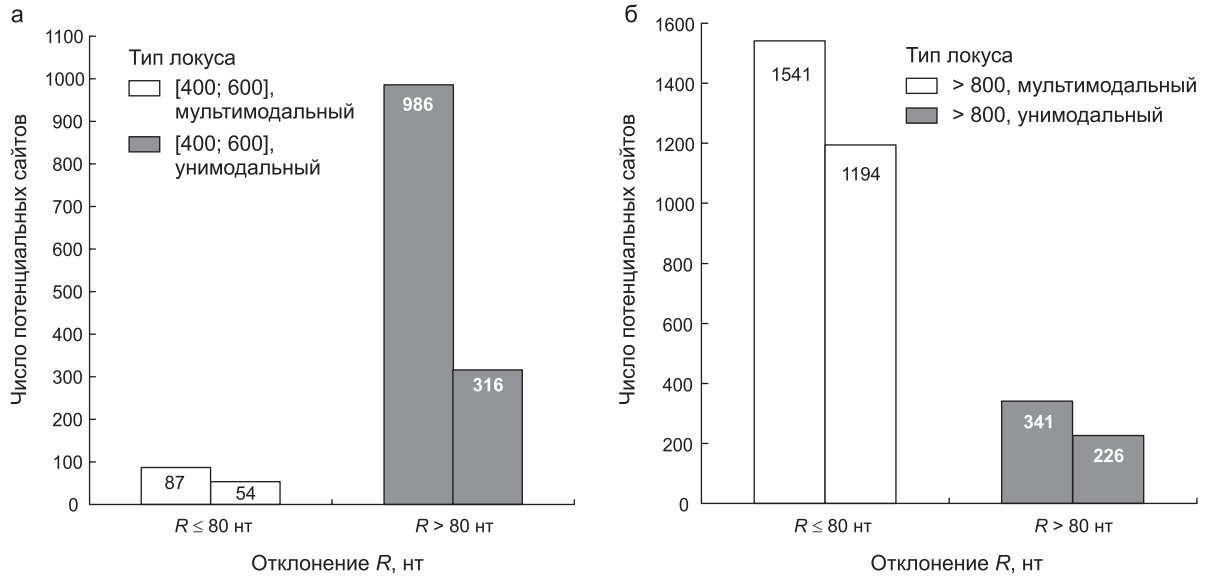

Рис. 8. Количество потенциальных сайтов, выявленных в выборках локусов диапазонов длин от 400 до 600 нт (а) и более 800 нт (б) и классифицированных согласно признакам: отклонение сайта от пика ≤ 80 нт или > 80 нт; тип локуса – уни- или мультимодальный.

по длине (табл. 2) и последующие расчеты плотностей близлежащих потенциальных сайтов позволяют более аккуратно интерпретировать обнаруженный нами ранее (рис. 3) эффект обогащения потенциальными сайтами районов вблизи пиков в унимодальных локусах по сравнению с мультимодальными локусами без учета длин локусов. Для этого следует учесть, что при расчете значимостей обогащения (рис. 4) в общую выборку унимодальных локусов входила выборка локусов длиной до 400 нт, для которых мы установили, что отклонение от пиков 100 % сайтов, выявленных в этих локусах, не превышает 180 нт. Мы предполагаем, что вследствие своего значительного объема именно эта группа дает наиболее заметный вклад в общую значимость, рассчитанную без учета длин локусов (рис. 3). Также ощутимый вклад в эту значимость дают различия пары выборок длиной от 400 до 600 нт и от 600 до 800 нт. Проведенный анализ показывает, что при длинах локусов свыше 800 нт нет ярко выраженного обогащения районов вблизи аннотированных пиков в унимодальных локусах по сравнению с мультимодальными (рис. 6). Учитывая эти соображения, а также принимая во внимание, что в случае присутствия единичного сайта по способу построения профилей локус не должен значительно превышать длину порядка 400 нт, можно предположить, что при длине свыше 600 нт локусы, определенные согласно нашей классификации как унимодальные, получены вследствие наличия не одного, а двух и более сайтов. Это позволяет заключить, что более корректной является следующая классификация локусов по форме пиков: 1) унимодальные локусы длиной до 600 нт, образованные предположительно одним сайтом связывания FoxA, и 2) мультимодальные локусы от 400 до 600 нт и все локусы большей длины, сформированные множеством отдельных сайтов. При этом мы считаем, что форма пика, на которой был основан наш способ классификации локусов, в случае длины локуса свыше 600 нт и наличия двух и более сайтов имеет стохастическую природу. Эта природа, в частности, зависит: 1) от расстояния между сайтами и 2) от аффинности каждого сайта, которая в ходе эксперимента ChIP-Seq интерпретируется как соответствующая высота пика. Поэтому локусы, отнесенные нами к первой группе, легче поддаются аннотации, более пригодны для использования методами, основанными на выявлении сходных контекстных мотивов для выявления сайтов связывания из данных ChIP-Seq. Локусы, отнесенные нами ко второй группе, предположительно, являются источниками неточностей при обработке данных ChIP-Seq, поскольку множество близко расположенных ССТФ создают более сложный паттерн покрытия, что может приводить к неточности локализации сайтов. Таким образом, настоящее исследование показывает, что при аннотации и обработке данных, полученных по технологии ChIP-Seq, необходимо принимать во внимание не только высоту пика профиля, но также его форму и протяженность локуса, что поможет устранить ряд неточностей при обработке этих данных.

Необходимо заметить, что в ходе данного исследования мы не ставили целью проведение строгой классификации локусов на два подмножества. Это связано также с тем, что есть определенный произвол в интерпретации конечных данных эксперимента ChIP-Seq. Например, а) варьирует длина фрагментов, на которые в ходе эксперимента разрезается геномная ДНК; б) выбранный нами согласно E. Wederell с соавт. (2008) порог максимальной высоты пика (10) влияет на анализ плотности сайтов с малым отклонением от пиков. Поэтому определенные нами здесь приблизительные границы подмножеств локусов применимы только к данным этого эксперимента ChIP-Seq (Wederell *et al*., 2008). Однако наличие такой классификации локусов может помочь в процедуре аннотации данных других ChIP-Seq экспериментов.

Работа выполнена при финансовой поддержке: Российского фонда фундаментальных исследований (грант № 09-04-00562-а); программы РАН 22.8 «Молекулярная и клеточная биология»; междисциплинарного интеграционного проекта СО РАН 119 «Постгеномная биоинформатика: компьютерный анализ и моделирование молекулярно-генетических систем»; госконтракта с ФАО П721 «Разработка программно-информационного комплекса для исследования механизмов регуляции экспрессии генов эукариот».

Литература

Левицкий В.Г., Ощепков Д.Ю., Ершов Н.И. и др. Разработка методов распознавания сайтов связывания транскрипционных факторов FoxA, их экспериментальная верификация и использование для анализа данных массовой иммунопреципитации хроматина // Докл. АН (в печати).

Dedon P.C., Soults J.A., Allis C.D., Gorovsky M.A. A simplified formaldehyde fixation and immunoprecipitation technique for studying protein-DNA interactions // Anal. Biochem. 1991. V. 197. P. 83–90.

Farnham P. Insights from genomic profiling of transcription factors // Nat. Rev. Genet. 2009. V. 10. P. 605–616.

Gershenzon N.I., Stormo G.D., Ioshikhes I.P. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites // Nucl. Acids Res. 2005. V. 33. № 7. P. 2290–2301.

Jothi R., Cuddapah S., Barski A. *et al*. Genome-wide identification of *in vivo* protein–DNA binding sites from ChIP-Seq data // Nucl. Acids Res. 2008. V. 36. № 16. P. 5221–5231.

Kolchanov N.A., Ignatieva E.V., Ananko E.A. *et al*. Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucl. Acids Res. 2002. V. 30. P. 312–317.

Levitsky V.G. Application of motif discovery tool for FoxA binding sites analysis // Proc. of the Seventh Intern. Conf. On Bioinformatics of Genome Regulation and Structure\System Biology (BGRS\SB'2010). 2010. P. 165.

Levitsky V.G., Ignatieva E.V., Ananko E.A. *et al*. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions // BMC Bioinformatics. 2007. V. 8. P. 481.

Robertson G., Hirst M., Bainbridge M. *et al*. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing // Nature Methods. 2007. V. 4. № 8. P. 651–657.

Wederell E.D., Bilenky M., Cullum R. *et al*. Global analysis of *in vivo* Foxa2-binding sites in mouse adult liver using massively parallel sequencing // Nucl. Acids Res. 2008. V. 36. № 14. P. 4549–4564.

Welsheimer T., Newbold J.E. A functional hepatocyte nuclear factor 3 binding site is a critical component of the duck hepatitis B virus major surface antigen promoter // J. Virol. 1996. V. 70. № 12. P. 8813–8820.

Zhang M., Marr T. A weighted array method for splicing and signal analysis // Comput. Appl. Biol. Sci. 1993. V. 9. P. 499–509.

ANALYSIS OF DATA OF LARGE-SCALE CHROMATIN IMMUNOPRECIPITATION BY METHODS OF PERCEPTION OF TRANSCRIPTION FACTOR BINDING SITES

V.G. Levitsky, G.V. Vasil'ev, D.Yu. Oshchepkov, N.I. Ershov, T.I. Merkulova

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia,
e-mail: levitsky@bionet.nsc.ru

Summary

Chromatin immunoprecipitation followed by massive parallel sequencing of the precipitated fragments (ChIP-Seq) is broadly used for detailed investigation of the distribution of various transcription factor binding sites over the whole-genome. The ChIP-Seq profiles obtained by immunoprecipitation of mouse liver chromatin with antibodies against the FoxA2 transcription factor (Wederell *et al*., 2008) were analyzed by our methods of recognition of FoxA binding sites. The following classification of locus profiles is proposed: (1) Unimodal loci (possessing a single peak) of lengths below 600 bp. These loci are likely to be formed by a single FoxA binding site. (2) Multimodal loci (possessing two or more distinct peaks) of lengths over 600 bp and all longer loci. Each locus of this group is likely to involve several sites.

Key words: FoxA transcription factors, binding sites, computer perception methods, chromatin immunoprecipitation data.