УДК 577.21:577.29:004.42

КОМПЬЮТЕРНОЕ ИССЛЕДОВАНИЕ РЕГУЛЯЦИИ ТРАНСКРИПЦИИ ГЕНОВ ЭУКАРИОТ С ПОМОЩЬЮ ДАННЫХ ЭКСПЕРИМЕНТОВ СЕКВЕНИРОВАНИЯ И ИММУНОПРЕЦИПИТАЦИИ ХРОМАТИНА

© 2014 г. **Ю.Л. Орлов**

Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия,
e-mail: orlov@bionet.nsc.ru

Поступила в редакцию 2 сентября 2013 г. Принята к публикации 1 февраля 2014 г.

Появление высокопроизводительных экспериментальных технологий секвенирования привело к бурному росту объема данных о структуре генома, распределении регуляторных районов генов в геноме, особенностях их взаимодействия.

Статья представляет обзор технологий, связанных с иммунопреципитацией хроматина: ChIP-PET, ChIP-seq, ChIA-PET. Описаны компьютерные методы анализа сайтов связывания транскрипционных факторов и структуры регуляторных районов в масштабе генома. Показаны подходы к решению задач, возникающих при аннотации геномных данных, определении сайтов связывания транскрипционных факторов и регуляторных районов.

Ключевые слова: секвенирование, иммунопреципитация хроматина, ChIP-chip, ChIP-seq, ChIP-PET, ChIA-PET, сайты связывания транскрипционных факторов, регуляция генной экспрессии.

СЕКВЕНИРОВАНИЕ ГЕНОМОВ И ИССЛЕДОВАНИЕ РЕГУЛЯЦИИ ТРАНСКРИПЦИИ

После секвенирования первых полных геномов в молекулярной генетике произошла технологическая революция, связанная с появлением экспрессионных микрочипов высокой плотности и технологий высокопроизводительного секвенирования ДНК. Становится возможной детальная полногеномная аннотация регуляторных геномных последовательностей по экспериментальным данным, полученных в результате массового параллельного секвенирования ДНК (Tucker *et al*., 2009). Аннотация в широком смысле заключается в сопоставлении функциональной информации районам генома человека. Кроме определения положения и структуры белок-кодирующих генов, аннотация включает описание некодирующей РНК, выделение регуляторных районов генов, исследование хромосомных аномалий и однонуклеотидных полиморфизмов, определение функции РНК и белков, предсказание вторичной и пространственной структуры белков (ENCODE Project Consortium, 2012). Важную роль в аннотации играют полногеномные методы исследования, основанные на секвенировании ДНК.

Исследование механизмов регуляции транскрипции генов – важная самостоятельная фундаментальная задача. Ключевую роль в контроле экспрессии генов на уровне транскрипции играют транскрипционные факторы – белки, специфически взаимодействующие с регуляторными районами генов и другими белками транскрипционной машины. Экспрессия гена в значительной степени определяется и другими условиями, к числу которых относятся: состояние хроматина в данном районе генома (открытый, закрытый); уровень метилирования ДНК; плотность нуклеосомной упаковки ДНК (Joseph *et al*., 2010). Компьютерные модели

транскрипции необходимы как для успешного предсказания особенностей экспрессии генов, так и для выполнения прикладных исследований, например реконструкции регуляторных сетей, создания генетических конструкций с заданными свойствами, исследования механизмов заболеваний, канцерогенеза, поиска мишеней для лекарств в токсикологических исследованиях, выявления ключевых биомаркеров (Tucker *et al*., 2009).

Реконструировать генные сети (регуляторные контуры) взаимодействующих генов транскрипционных факторов можно через определение их регуляторных последовательностей в промоторах генов – мишеней их воздействия. Регуляторные районы групп коэкспрессирующихся генов зачастую имеют общие черты организации, что выражается в наличии регуляторных паттернов (цис-регуляторных модулей), состоящих из устойчивых сочетаний сайтов связывания транскрипционных факторов различных типов и других мотивов (Chen *et al*., 2008). Выявление и анализ таких регуляторных паттернов являются основой для построения обобщенных моделей регуляторных районов группы коэкспрессирующихся генов.

В настоящее время накоплен колоссальный объем данных в области регуляции экспрессии генов эукариот, наблюдается их непрерывный рост как первичных данных, доступных, в частности, в репозитории данных GEO NCBI (Gene Expression Omnibus) (www.ncbi.nlm.nih.gov/geo/). Все большую актуальность приобретают формализация описания механизмов регуляции транскрипции и разработка на этой основе методов интеграции гетерогенной информации об особенностях регуляции экспрессии генов, использующих как технологии определения сайтов связывания транскрипционных факторов в геноме, так и оценки уровней экспрессии генов на микрочипах.

Начавшееся в постгеномную эру (период после секвенирования первого «чернового» варианта генома человека в 2001 г.) соревнование компьютерных и экспериментальных технологий принято называть «геном за 1000 долларов» (Kedes, Campany, 2011) (Archon X Prize, см. http://genomics.xprize.org). Такое название произошло от объявленной в качестве цели цены за полное секвенирование генома человека. Интересно отметить, что соревнование производителей технологий секвенирования было настолько эффективным, что условия конкурса усложнялись несколько раз, и в итоге в конце 2013 г. он был впервые в истории отменен, остановившись на достигнутой отметке в 3–5 тыс. долларов за полное секвенирование генома человека. Не останавливаясь на названиях компаний и технологий, отметим, что доступность секвенирования, помимо огромного вклада в фундаментальные научные исследования, позволяет использовать секвенирование индивидуального генома как стандартный диагностический тест, что исключительно важно для медицинских приложений.

Отметим основные технологии высокопроизводительного секвенирования (Kedes, Campany, 2011): параллельное пиросеквенирование на микробусах, технология Roche 454, технология Illumina Solexa (http://www.illumina.com), использующая оптическое сканирование флюоресценции меченых нуклеотидов в клонированных колониях молекул ДНК на твердой поверхности, и секвенирование с помощью лигирования SOLiD (Sequencing by Oligonucleotide Ligation) компании «Applied Biosystems» (http://www.appliedbiosystems.com). Перспективны новые технологии Ion Torrent, использующие детекцию ионов водорода во время полимеризации ДНК на гиперчувствительном сенсоре. Сама компания, внедрившая эту технологию секвенирования, была поглощена компанией «Life Technologies», которая в свою очередь поглощена компанией «ThermoFisher» (www.thermofisher.com).

Секвенирование ДНК на наносферах (nanoball sequencing) компании «Complete Genomics» (www.completegenomics.com) использует циклическую амплификацию фрагментов геномной ДНК по принципу «катящегося кольца». Заметим, что в 2013 г. «Complete Genomics» уже поглощена Пекинским институтом биоинформатики (BGI), что свидетельствуют о выходе Китая на лидирующие позиции в мире в технологиях секвенирования. Компания «Pacific Biosciences» (PacBio) предлагает альтернативную технологию определения последовательности одиночной молекулы ДНК (технология SMRT) при считывании ДНК-полимеразой (www.pacificbiosciences.com).

Общий тренд в сравнении представленных технологий: чем ниже цена за секвенирование за нуклеотид (за мегабазу) и выше производительность технологии секвенирования, тем короче получающиеся фрагменты ДНК. Секвенирование по технологии 454 позволяет получать последовательности до 300 нуклеотидов, в то время как следующие технологии, такие как SOLiD, – не более 50–100 нуклеотидов. Повышение производительности секвенирования ставит технически более сложные задачи биоинформационного анализа таких данных, выравнивания, картирования и ассемблирования коротких последовательностей ДНК.

ТЕХНОЛОГИЯ ИММУНОПРЕЦИПИТАЦИИ ХРОМАТИНА

Среди методов, позволяющих изучать связывание транскрипционных факторов (ТФ) с ДНК, наибольшими перспективами для полногеномных исследований обладает метод иммунопреципитации хроматина (Chromatin ImmunoPrecipitation – ChIP).

Процедура ChIP-seq на первом этапе требует фиксации контактирующих молекул ДНК и белков в клетке с помощью формальдегида, что вызывает образование ковалентных сшивок между ДНК и белками. Затем хроматин дробится ультразвуком на фрагменты (существует термин «соникация», или дословно, «озвучивание» хроматина). Альтернативно разделение ДНК может выполняться с помощью разрезания ферментами рестрикции. Затем с помощью иммунопреципитации со специфическими антителами выделяется ДНК, с которой физически связан интересующий исследователя белок. Белковая фракция отмывается, а ДНК (относительно короткие фрагменты не более нескольких сотен оснований) направляется на секвенирование с помощью имеющегося оборудования массового параллельного секвенирования ДНК (Roche 454, Illumina или SOLiD).

Прочитывание ДНК выполняется не для всей последовательности экстрагированного фрагмента, составляющей до нескольких сотен нуклеотидов, а для крайнего участка – от 20 до 75 п.о. Если выполняется лигирование концов фрагмента ДНК, они могут секвенироваться попарно (так называемые парные концы, или PET – Paired End Tags). На рис. 1 представлена схема определения кластеров фрагментов ДНК на хромосоме для парных фрагментов (парных концов метода ChIP-PET) и для одиночных фрагментов. Задача картирования секвенированных фрагментов (прочтений, или «ридов», reads) технически выполняется с помощью выравнивания

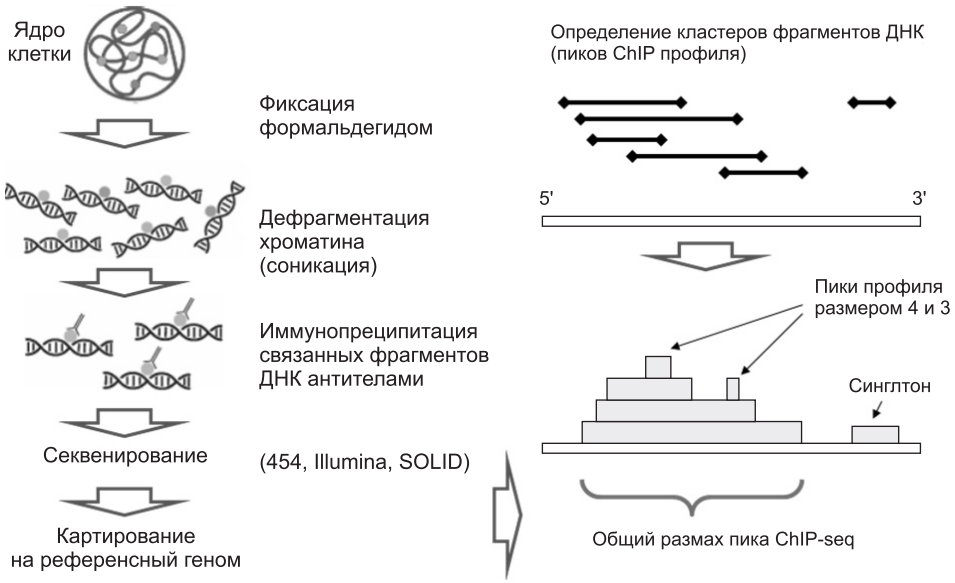

Рис. 1. Схема иммунопреципитации хроматина, картирования фрагментов ДНК и пример получаемого ступенчатого профиля ChIP-seq.

и быстрого поиска совпадений с последовательностями ДНК хромосом генома.

Картирование секвенированных последовательностей технически требует до нескольких часов машинного времени на персональном компьютере для достаточно больших по размеру (порядка 3Гб) геномов эукариот, таких как геном мыши. Для решения этой задачи применяются компьютерные методы оптимизации выравнивания и быстрого поиска совпадений в нуклеотидных последовательностях. Существует набор программ картирования прочтений для форматов данных Illumina и форматов цветовой кодировки SOLiD (см. обзор программ на SEQanswers, http://seqanswers.com). Далее с использованием координат секвенированных фрагментов на хромосомах референсного генома строится так называемый геномный профиль и выполняется его анализ (рис. 2).

При построении профиля связывания координаты прочтений удлиняются до размера исходного фрагмента ДНК (150–200 п.о.) и получается «лестница» или «сток» фрагментов, наложенных друг на друга в геномных координатах. При использовании парных концов длина фрагментов известна и удлинение не нужно. Затем определяются пики такого геномного профиля. Пик – наиболее высокая точка локального профиля. Пик более вероятно содержит сайт связывания транскрипционного

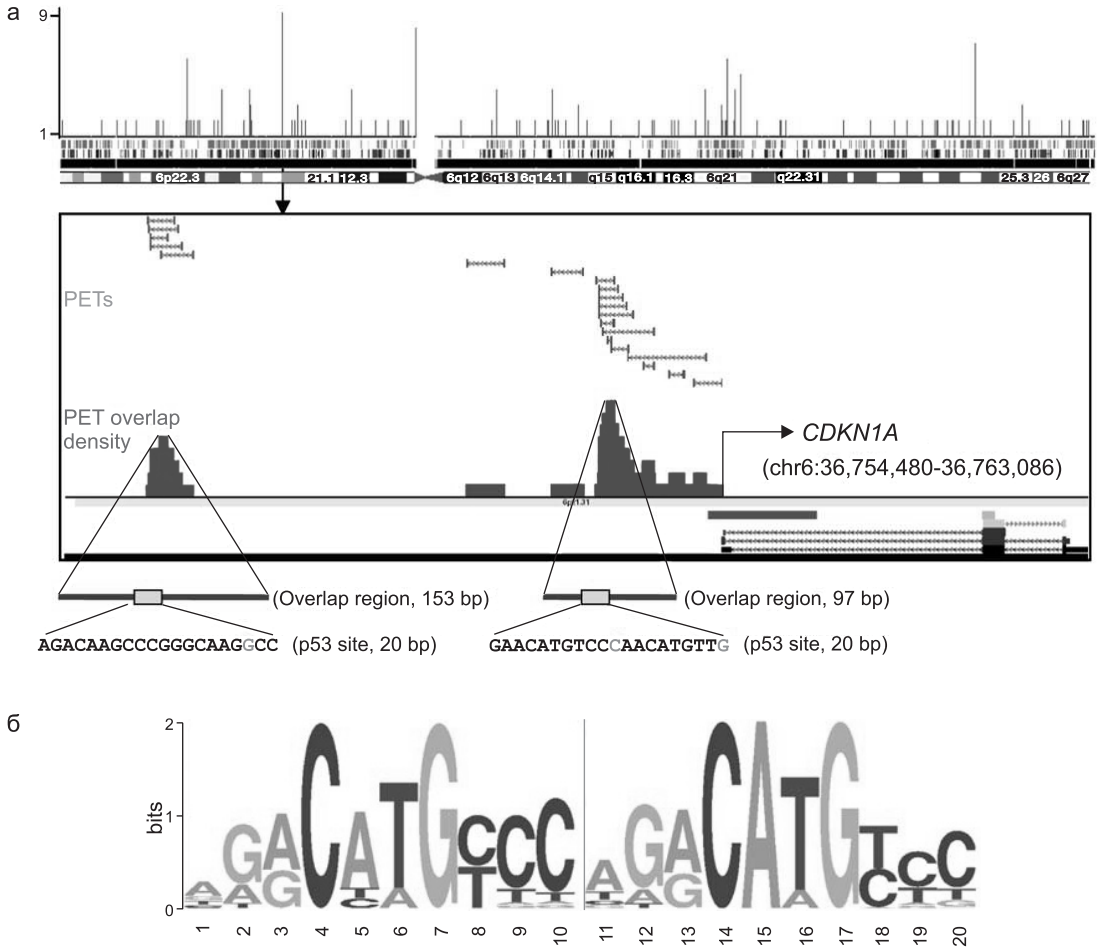

Рис. 2. Полногеномное представление сайтов связывания транскрипционного фактора p53 на хромосоме 6 человека (а) и выделенный стрелкой район генома, содержащий ген *CDKN1A* и его 5′-район в большем масштабе.

Два участка профиля, сформированного фрагментами ДНК, прошедшими иммунопреципитацию, образуют пики, содержащие узнаваемый консенсусный мотив связывания p53 · 5′-AGA**CATG**CCC**AGACATG**CCC-3′. На панели (б) показана частотная матрица мотива связывания, восстановленная по полногеномным данным (Wei *et al*., 2006. P. 207–219).

фактора, чем окружающий геномный район, поскольку все фрагменты ДНК, составляющие «ступени» пика, были связаны с белком в ChIP эксперименте.

Исторически метод иммунопреципитации хроматина предназначался для анализа взаимодействий белок–ДНК на единичной или ограниченной выборке данных (несколько промоторных участков). Массовое использование олигонуклеотидных микрочипов позволило усовершенствовать технологию и получать гибридизационный сигнал связывания исследуемой последовательности с заранее подготовленными пробами. Технология получила название ChIP-on-chip, или ChIP-chip (т. е. хроматин-иммунопреципитация на микрочипе) (Collas, Dahl, 2008). Такой анализ возможен для достаточно больших наборов проб, включая, например, отдельные хромосомы или все промоторные районы генов, известные в геноме. Существуют варианты микрочиповых технологий для определения связывания с белками в специализированных вариантах: метод DamID (использующий белок Dam – DNA adenine methyltransferase), метод DIP-seq (DNA ImmunoPrecipitation) для исследования связывания с ДНК без хроматина (Liu *et al*., 2005).

Общий набор методов выделения сайтов связывания, основанных на иммунопреципитации, по-английски называют «chop and chip» или «Chop and ChIP», что можно перевести как «нарубить и расщепить» или «разделить на (микро)чип» (Collas, Dahl, 2008). Технология ChIP-seq применялась для поиска сайтов связывания ТФ мыши (Chen *et al*., 2008), человека (Wei *et al*., 2006; Zeller *et al*., 2006; Chia *et al*., 2010), анализа модификаций гистонов в различных тканях (Joseph *et al*., 2010).

Исследование профиля ChIP-seq требует специализированных компьютерных программ и ориентированных на конкретную задачу методов, в зависимости от технологий секвенирования (коррекция на специфические ошибки), размера и особенностей генома (наличие повторенных последовательностей, детали аннотации – от компактного, хорошо изученного генома дрожжей до большого по размеру генома человека и, например, геномов растений, где нет референсной последовательности). Первичные данные секвенирования поступают в формате bed-файлов либо в FASTA формате для картирования на геном (тысячи и миллионы последовательностей).

В результате стандартного эксперимента ChIP-seq (Chen *et al*., 2008) по извлечению белка, специфически связанного с ДНК, получается набор последовательностей, упорядоченное расположение которых в геноме (в хромосомных координатах) дает ступенчатый профиль. Короткие секвенированные фрагменты уже на геномной карте удлиняют до 150–300 нуклеотидов, что в среднем соответствует размеру экстрагированной в ChIP-seq ДНК (и соответствует размеру 1–2 нуклеосом). Заметим, что практически достаточно секвенировать первые 25–35 нуклеотидов, чтобы однозначно найти координаты фрагмента в геноме.

Образующийся численный профиль связывания вдоль хромосомы (одномерная координата) содержит интересующие исследователя места в геноме – участки (сайты) связывания с белком. При увеличении масштаба в геномном браузере видны пики (горки) профиля (рис. 2). Пики получаются в результате перекрывания нескольких фрагментов (графически как наложенные друг на друга полоски в одномерных координатах хромосомы). Пик профиля как сигнал можно математически описать координатами начала и конца участка, указать высоту пика, положение вершины (саммита), оценить вероятность получения пика по случайным причинам. При близком расположении нескольких участков, связанных с исследуемым белком, пик может иметь несколько вершин (мультимодальный пик), что задает неоднозначность при выборе позиции специфичного сайта.

Секвенирование фрагментов ДНК может выполняться с двух сторон, с использованием технологии секвенирования парных концов, что позволяет более точно картировать сайты. Отметим, что относительно недавние работы 2005–2007 гг., в которых использовалось секвенирование парных концов (PET – Paired End Tags в аббревиатуре ChIP-PET), включали клонирование выделенных фрагментов ДНК как необходимый технологический шаг перед секвенированием (Wei *et al*., 2006; Zeller *et al*., 2006). Клонирование увеличивало общее время эксперимента и вело к частичной потере данных (при неравномерном клонировании фрагментов ДНК по длине). В настоящее время метод ChIP-

seq, основанный на прямом секвенировании полученных после иммунопреципитации фрагментов ДНК, включает возможность использования протокола секвенирования парных концов.

В целом секвенирование обладает рядом преимуществ по сравнению с микрочиповыми технологиями исследования связывания белков в регуляторных районах ChIP-on-chip. Во-первых, результатом ChIP-seq является картина полногеномного распределения сайтов связывания транскрипционных факторов (ССТФ). Во-вторых, полученный результат свободен от предварительной селекции исследуемых районов, которые могут существенно искажать конечный результат. В-третьих, результатом ChIP-seq являются не только относительные уровни сигнала связывания, а конкретные участки последовательностей ДНК, что дает больше возможностей для теоретического анализа и точного определения сайтов.

Метод иммунопреципитации хроматина с последующим секвенированием всего пула выделенных фрагментов широко используется для получения картины полногеномного распределения сайтов связывания различных транскрипционных факторов (Chen *et al*., 2008; Chia *et al*., 2010). Более 160 профилей связывания для генома человека представлены в данных проекта ENCODE (ENCODE Project Consortium, 2012), десятки профилей ChIP-seq доступны для генома мыши, других модельных организмов.

МАТЕМАТИЧЕСКАЯ ЗАДАЧА АНАЛИЗА ПРОФИЛЯ СЕКВЕНИРОВАНИЯ

Приведем немного цифр. В типичной задаче анализа профиля ChIP-seq геном человека (размер около 3 Гб) разделен по хромосомам от 20 до 200 Мб. Профили строятся для каждой хромосомы. Стандартно каждой позиции нуклеотида ставится в соответствие высота профиля, тем самым линейно увеличивая размер файла в зависимости от размера генома. Однако при более компактной записи профиля размер файла чуть меньше, порядка 100 Мб. При дополнительной аннотации размеры файлов значительно увеличиваются, фактически на пределе возможностей работы стандартного персонального компьютера. Для других организмов геном может иметь как меньший (дрожжи, нематода), так и существенно больший размер (геномы растений, таких как пшеница); практически объем данных ChIP-seq составляет от 100 Мб до 1 Гб.

В результате секвенирования ChIP-seq имеем набор коротких фрагментов (прочтений, или reads, tags, размер от 25 до 50 нуклеотидов). Эти фрагменты распределены по геному неравномерно. Размер библиотек (наборов данных секвенирования, получаемых в одном эксперименте) колеблется от 2 до 20 млн коротких последовательностей. Фиксируя одну из хромосом, рассматриваем положение последовательностей на ней, определяем одномерные координаты, ищем сайты связывания белка на ДНК. Сайт в такой формулировке – небольшой участок, 6–8 нуклеотидов. Для анализа мотива связывания достаточно определить координаты его центра (см. рис. 3) и сравнить с известными базами данных мотивов связывания (Laajala *et al*., 2009).

Уникальность (однозначность) картирования короткой последовательности на геном представляет особую проблему анализа данных. Если в геноме был повтор (два участка ДНК в разных местах генома, достаточно длинных, скажем, 100 п.о. и более) и наш короткий фрагмент ДНК попал в повтор, то мы не можем его однозначно (уникально) картировать. Пример таких затруднений – картирование регуляторных сайтов для генов, имеющих псевдогены. Общее свойство протяженных геномных последовательностей – иметь участки, недоступные для однозначного картирования короткими фрагментами из-за длинных совершенных повторов. Возникли термины «картируемость» (mappability) и «уникома» (uniqueome) – как часть генома, которая однозначно (уникально) определяется короткими фрагментами заданной длины (Lee, Schatz, 2012). Для каждой длины секвенированных ридов существует своя «уникома». Например, для фрагментов размером 50 п.о. некартируемых участков гораздо меньше, чем для фрагментов 25 п.о. Наборы таких карт можно рассчитать, существуют и готовые разметки «уникальности» для нескольких референсных геномов, в частности геномов человека и мыши (Lee, Schatz, 2012). Отметим, что программы

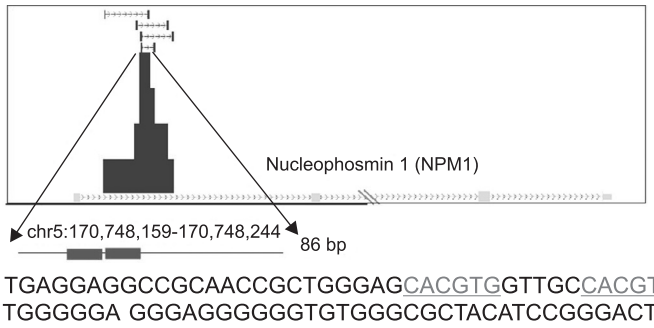

Рис. 3. Расположение секвенированных парных концов ChIP-PET и анализ нуклеотидной последовательности связывания Myc.

Известный сайт связывания Myc в первом интроне гена *NPM1* определен кластером PET-4. Показаны перекрывающиеся PET фрагменты (вверху), образующийся ступенчатый профиль (пик) пересечения фрагментов, координаты в геноме человека. Внизу показана последовательность, соответствующая пересечению всех фрагментов ДНК в кластере с выделенными элементами, соответствующими каноническому E-боксу (Zeller *et al*., 2006. P. 17834–17839).

разметки «уникальности» требуют высокопроизводительных компьютерных вычислений – фактически, поиска всех повторов в геноме.

Поскольку эксперимент ChIP-seq проводится на пуле (большом наборе) клеток, каждый акт связывания, представленный фрагментом ДНК, не зависит от других последовательностей. Перекрывание фрагментов в геноме в районе пика подтверждает специфичность связывания. Истинные (специфичные) пики профиля имеют бóльшую высоту (сформированы бóльшим числом фрагментов ДНК), что можно показать статистически. Для определения порогового значения пика, когда с уверенностью можно говорить про специфичность, удобно упорядочить все полученные пики по высоте и рассчитать число пиков в зависимости от высоты (рис. 4). Полученное распределение всегда очень неравномерно – большинство пиков имеет небольшую высоту, затем число высоких пиков быстро (почти экспоненциально) убывает.

Задача определения набора пиков в геноме решается статистически с помощью сравнения экспериментально полученного распределения

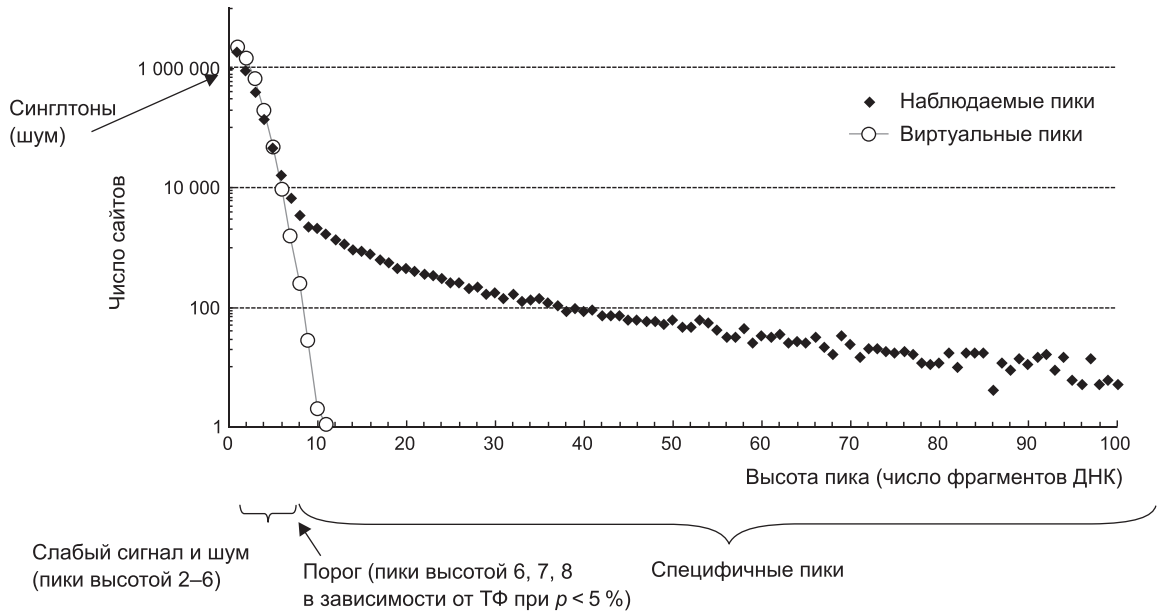

Рис. 4. Распределение числа сайтов в геноме в зависимости от высоты пика ChIP-seq (для связывания ТФ Nanog мыши) (Chen *et al*., 2008. P. 1106–1117).

набора пиков к ожидаемому по случайным причинам (рис. 4). Был предложен компьютерный алгоритм определения пиков профиля и их последующей фильтрации от «шума» и ошибок секвенирования, основанный на статистике распределения числа сайтов в геноме в зависимости от высоты пика ChIP-seq и сравнении специфического профиля с контрольным. Контрольное секвенирование выполняется после разделения геномной ДНК ультразвуком без антител или с иммунопреципитацией к неспецифическому белку, например GFP или IgG. Такой подход в разных вариантах реализован в компьютерных программах определения пиков профиля ChIP-seq (Zhang *et al*., 2008).

Распределение вероятности наблюдения $P_{obs}(X = m)$ сайтов (пиков профиля) X фиксированной высоты m может быть представлено как взвешенная сумма специфического и неспецифического распределений

$$P_{obs}(X = m) = \\ = \alpha \cdot P_{sp}(X = m) + (1 - \alpha) \cdot P_{ns}(X = m), \quad (1)$$

где P_{obs} – функция вероятности наблюдаемого распределения встречаемости ChIP пиков, X – пики, m = 1, 2, 3, ... – высота пика, P_{sp} – вероятность специфичных пиков, $0 < \alpha < 1$ – доля специфичных пиков в общем распределении пиков в геноме, P_{ns} – вероятность неспецифичных пиков (шумовой сигнал) в общем распределении пиков по профилю в геноме (Kuznetsov *et al*., 2007).

На основе экспериментальных ChIP данных мы можем построить эмпирическую функцию распределения пиков в геномном профиле. Распределение числа специфичных пиков P_{sp} в зависимости от их высоты может быть промоделировано с помощью распределения Пуассона или распределения Парето, начиная с большой высоты пика, где шумового сигнала уже нет (например, с m = 10 на рис. 3) (Kuznetsov *et al*., 2007). Неспецифичное распределение P_{ns} может быть оценено с помощью компьютерной симуляции по появлению кластеров (пиков) при случайном (равномерном) распределении прочтений (секвенированных фрагментов ДНК) вдоль хромосомы. Разработанная компьютерная модель симуляции случайных пиков в геноме требует достаточно интенсивных пересчетов (несколько часов работы на персональном компьютере). Позднее в качестве контроля использовалось число пиков неспецифичного секвенирования (без иммунопреципитации) всей доступной ДНК или с использованием неспецифичных белков (таких как GFP, IgG). Специфичность связывания P_{sp} зависит от силы связывания ДНК с белком, что может быть подтверждено далее с помощью выборочного тестирования методом qPCR (Joseph *et al*., 2010).

Для опубликованных ТФ в геноме получается от 1000 до 20 000 сайтов связывания, подтвержденных (выборочно) независимым тестированием (Chen *et al*., 2008; Joseph *et al*., 2010). Стандартно получают около 5 тыс. сайтов (мест на хромосоме), каждый со своей высотой пика, характеризующей его «силу связывания». Сила связывания определяется аффинностью короткой последовательности ДНК – теми самыми 6–8 нуклеотидами мотива связывания ТФ. Показано, что высота пика профиля ChIP-seq в геномном эксперименте коррелирует со связыванием в отдельных проверочных экспериментах qPCR.

Истинные пики геномного профиля, как правило, содержат мотив сайта связывания – последовательность, похожую на известную консенсусную последовательность ДНК – стандарт, известный из других экспериментов (рис. 2). Мотив обычно выражается не одним стандартом (консенсусом), а допускает несовпадения в нуклеотидной последовательности, описываемые позиционной весовой матрицей для данного сайта. Такие весовые матрицы для многих транскрипционных факторов есть в базах данных, таких, как TRANSFAC, TRRD (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/), JASPAR (http://jaspar.genereg.net/) (результаты «одиночных» разрозненных экспериментов или более старых технологий, собранные по литературе). Высокий процент содержания в пиках ChIP-seq исследуемого мотива, оцененного весовой матрицей, показывает, что положение пиков определено правильно.

Другая оценка качества сигнала связывания в профиле ChIP-seq – это относительная высота пика по отношению к контролю – отношение числа специфических фрагментов ДНК в исследуемой точке генома к числу неспецифических, полученному в контрольном эксперименте секвенирования без иммунопреципитации (Bailey *et al*., 2013). Контрольный профиль получают в

результате эксперимента по секвенированию в тех же условиях, но без специфических белков («пустой» прогон). Обычно требуется превышение высоты пика по отношению к контролю в 3–5 раз. Используя контрольный профиль секвенирования, можно получить оценку ошибки ложного предсказания для каждого пика, FDR (False Discovery Rate).

Есть ряд опубликованных программ анализа пиков ChIP-seq, например GLITR (GLobal Identifier of Target Regions), MACS (Zhang *et al*., 2008), HPeak, PeakFinder, QuEST, CisGenome, USeq и PICS (см. для обзора Laajala *et al*., 2009; Bailey *et al*., 2013).

Для картирования длинных фрагментов ДНК хорошо подходят программы MUMmer и BLAT. Для картирования коротких фрагментов ДНК набор программ достаточно велик: MAQ (Mapping and Assembly with Quality), SOAP (Short Oligonucleotide Alignment) (Li *et al*., 2008), использующая индексы для представления референсной последовательности; программа ELAND, ориентированная на стандарт данных Illumina (http://www.illumina.com/systems.ilmn).

Распространены также программы SeqMap, RMAP, ZOOM (Bailey *et al*., 2013), Bowtie, использующая индекс Burrows–Wheeler (BWT) (Langmead *et al*., 2009). Сегодня необходимо развитие методов для более широкого их использования для геномов с недостаточной аннотацией.

В целом задачи компьютерной обработки полногеномных данных секвенирования можно разделить на следующие направления:

– первичная фильтрация данных, картирование;

– анализ профилей секвенирования ДНК, сопряженного с иммунопреципитацией (ChIP-seq), с выделением как точечных участков связывания, так и протяженных участков (модификации гистонов);

– разметка сайтов связывания нуклеосом по нуклеотидной последовательности;

– интеграция данных секвенирования между собой (кластеры сайтов связывания транскрипционных факторов);

– интеграция данных с геномной аннотацией (определение генов-мишеней).

Программа HPeak (Hidden Markov model Peak) использует скрытые марковские модели для определения геномных участков, контактирующих с белками. Используется статистическая модель, учитывающая реалистичное распределение вероятностей для прочтений ДНК.

Программа MACS (Zhang *et al*., 2008) исходно была ориентирована на технологию Illumina Solexa (данные Solexa Genome Analyzer). Программа использует фрагменты («риды») в противоположных ориентациях, чтобы определить так называемый размер сдвига – близость между «ридами», содержащими сайты связывания. Преимуществом MACS является локальное моделирование «шумового», или контрольного, секвенирования с помощью распределения Пуассона по участкам хромосом (участков, размер которых задается пользователем, 1–10 т.п.о.).

Рассмотрим пример серии экспериментов ChIP-seq для одного и того же типа клеток. Масштабное исследование профилей связывания 13 различных транскрипционных факторов (ТФ) в геноме мыши было выполнено в работе (Chen *et al*., 2008) на эмбриональных стволовых клетках. Исследовались ТФ, значимые для поддержания плюрипотентности и развития организма, такие как Oct4, Nanog, Sox2. Использовалась платформа секвенирования Illumina Solexa (рис. 5). Для каждого транскрипционного фактора было получено от 5 до 12 млн последовательностей, в среднем 65 % из них картировались однозначно на референсный геном мыши, что соответствует стандартам экспериментов ChIP-seq. Для контроля использовалась иммунопреципитация к неспецифическому белку GFP (Green Fluorescent Protein). Определение значимых кластеров (пиков) профиля выполнялось с помощью сравнения профилей с контрольным секвенированием и последующей нормализации. Обычно получается несколько тысяч сайтов на геном для одного транскрипционного факторов. В работе X. Chen с соавт. (2008) было определено от 1,126 до 39,609 сайтов связывания для этих 13 ТФ.

При «зашумлении» данных или недостаточно глубоком секвенировании возникают специфические статистические проблемы. Например, при неглубоком секвенировании (0,5–1 млн ридов вместо 5–10 млн) геном человека не покрыт полностью секвенированными фрагментами и трудно гарантировать, что все сайты в эксперименте выявлены. Ранее считалось (эмпирически), что 5–6 млн фрагментов уже

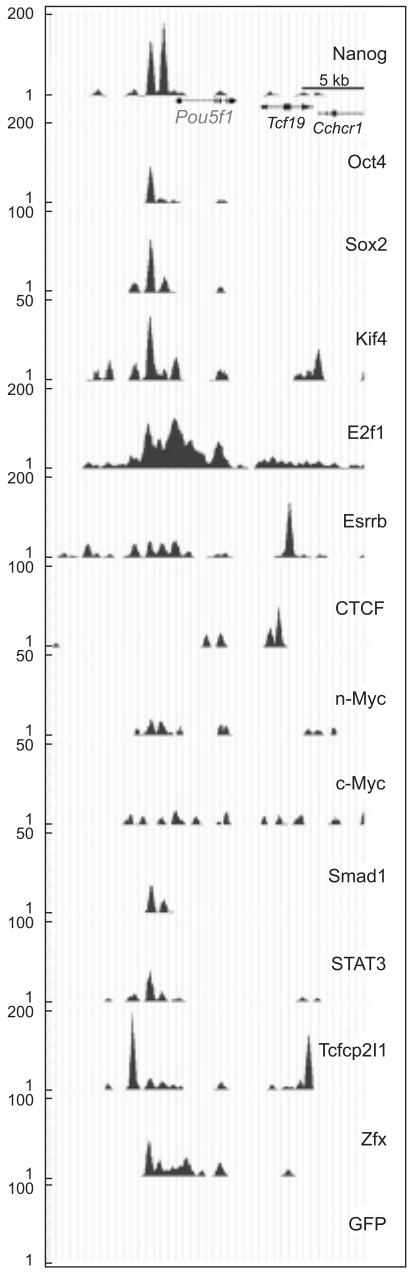

Рис. 5. Профили связывания 13 различных транскрипционных факторов в геноме мыши для гена *Pou5f1*.

Внизу представлен профиль контрольного секвенирования (для неспецифического белка GFP) (Chen *et al*., 2008. P. 1106–1117).

достаточно, чтобы выявить все сайты в геноме (Chen *et al*., 2008). Сейчас стандарт эксперимента ChIP-seq – это 10–20 млн и выше прочтений ДНК. Можно показать статистически, что при дальнейшем увеличении глубины секвенирования просто повышается порог высоты пика при выделении пиков из профиля и уже не детектируются новые сайты в геноме (в данных условиях эксперимента – при фиксированном антителе), что обосновывает достаточность использования меньшего числа прочтений (Chen *et al*., 2008).

Следует отметить, что описанные возможности применимы для хорошо изученных организмов, геномы которых не только известны, но и хорошо аннотированы (человек, мышь, *A. thaliana*, нематода *C. elegance* и др.). Однако существует большое число важных модельных организмов, геномы которых могут быть плохо аннотированы (рыба *D. rerio*) или недостаточно исследованы (слабо изученные геномы, например, паразитические черви). Тем не менее для модельных медицинских и биотехнологических задач изучение этих организмов очень актуально и эксперименты ChIP-seq выполнимы.

Полногеномное секвенирование коротких фрагментов без специфического белка может использоваться и для определения нуклеосомной упаковки (позиционирования всех нуклеосом) в полном геноме эукариот (Kaplan *et al*., 2009). Исследовалась, в частности, нуклеосомная упаковка в геноме дрожжей и ее взаимодействие с сайтами связывания транскрипционных факторов. При этом данные секвенирования ДНК (прямое секвенирование нуклеосомных фрагментов без иммунопреципитации) интегрировались с данными о сайтах связывания, определенными первоначально посредством ChIP-chip технологии (Kaplan *et al*., 2009).

СЕКВЕНИРОВАНИЕ И ТРЕХМЕРНАЯ СТРУКТУРА ХРОМОСОМ

Проблема анализа трехмерной структуры генома активно исследуется различными методами, использующими геномное секвенирование (рис. 6). Интересно отметить технологии Hi-C (Dekker *et al*., 2013) и ChIA-PET (Fullwood *et al*., 2009), позволяющие получать информацию о пространственной структуре хромосом через секвенирование. Метод Hi-C (Dekker *et al*., 2013) позволяет реконструировать карту пространственных взаимодействий хромосом в ядре клетки без специфических взаимодействий, не используя иммунопреципитацию.

Метод ChIA-PET (Chromatin Immunoprecipitation Analysis – Paired End Tags) (Fullwood *et*

al., 2009; Li *et al*., 2012), использующий иммунопреципитацию хроматина, позволяет определять взаимодействующие участки хромосом, контакты которых опосредованы белками или белковыми комплексами (рис. 6).

Для белка ERα – рецептора эстрогенов – выполнялся анализ карт контактов на хромосомах по методу ChIA-PET (Fullwood *et al*., 2009) (рис. 6). Данные хромосомных контактов, полученные с помощью полногеномного секвенирования, независимо подтверждались с помощью экспериментов по технологии 3C (Chromosome conformation capture) и флуоресцентной гибридизации *in situ* (FISH).

Эксперименты ChIA-PET по определению хромосомных контактов в ядре клетки, опосредованных уже не отдельным транскрипционным фактором, а целым транскрипционным комплексом РНК-полимеразы II, представлены в работе (Li *et al*., 2012). Был проведен компьютерный анализ распределения хромосомных контактов относительно генов и относительно участков генома, ассоциированных с модификациями хроматина (модификации гистона H3, связанные с открытым состоянием хроматина).

Предложена классификация моделей промоторных, энхансерных и мультигенных контактов, опосредованных комплексом РНК- полимеразы II (Li *et al*., 2012), включающая классы: базальный промотор, промотор-энхансер и мультигенная модель (см. рис. 7). Модель базального промотора включает только локальные петли ДНК в промоторе, без удаленных взаимодействий. Модель одиночного гена включает только петли в районе гена – между энхансером и промотором, возможно, между 5′- и 3′-районами гена, но без других белок-кодирующих генов. И наконец, мультигенная модель включает сразу несколько генов, расположенных рядом друг с другом и контактирующих промоторными районами. Введен термин «хромоперон» (chromoperon), или «хромосомный оперон». В этой модели также возможен контакт промоторов с удаленными энхансерами.

Для связывания транскрипционных факторов в геноме важно состояние хроматина, его открытость (отсутствие нуклеосомной упаков-

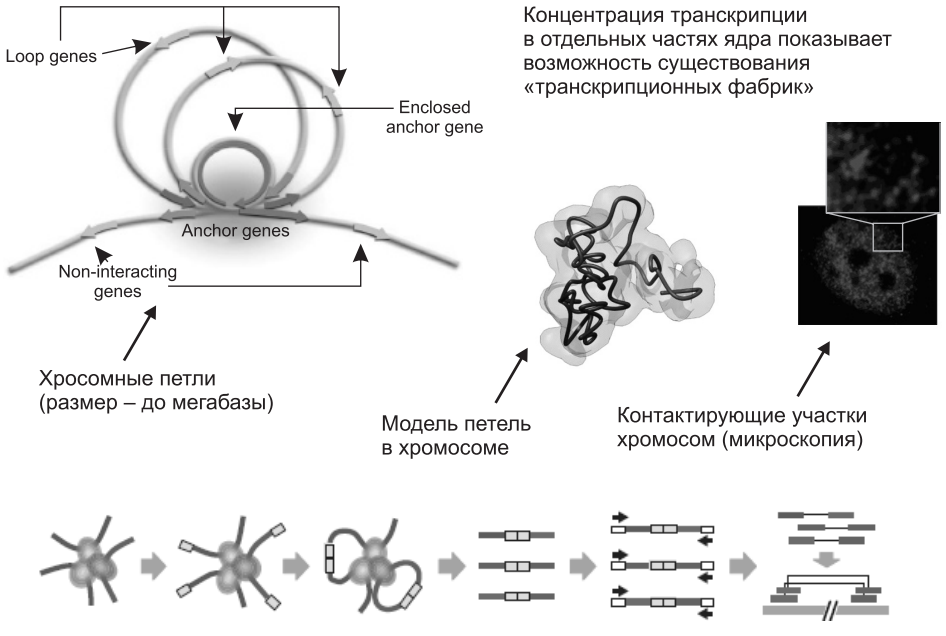

Рис. 6. Определение петель хромосом методом ChIA-PET.

а – общая схема структуры петель хромосом. Представлены контактирующие участки хромосом – петли и домены, определяемые методом ChIA-PET (Fullwood *et al*., 2009); б – схема метода ChIA-PET. Фрагменты ДНК из разделенных ультразвуком, прошедших иммунопреципитацию хроматиновых комплексов обрабатываются через лигирование линкеров (на свободные концы ДНК), лигирование сближенных фрагментов, получаются парные концы (PET). Далее ДНК секвенируется и картируется на референсную последовательность генома.

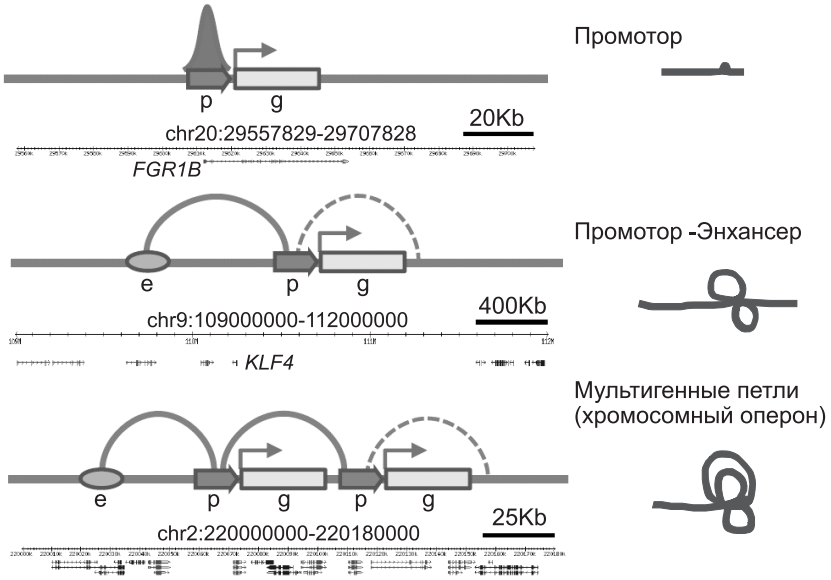

Рис. 7. ChIA-PET. Модели промоторных, энхансерных и мультигенных контактов, опосредованных комплексом РНК-полимеразы II (Li *et al*., 2012. P. 84–98).

ки), доступность ДНК. Маркеры модификаций гистонов, прежде всего гистона H3, модификации лизина в позициях 4, 14, 36, включающие метилирование и ацетилирование, связаны с доступностью ДНК и могут быть определены экспериментально также по технологии ChIP-seq (Joseph *et al*., 2010). Связь хромосомных контактов и модификаций хроматина (метилирование и ацетилирование) гистонов (гистона H3) исследовалась статистически в масштабе генома (Li *et al*., 2012). На рис. 8 показаны профили модификаций хроматина для участка хромосомы 2 человека, содержащей ген *PCBP1* вместе с профилями других ChIP экспериментов.

Исследование контактов хромосом, опосредованных комплексом РНК-полимеразы II в культурах клеток человека, подтвердило ассоциацию с сайтами связывания факторов транскрипции и модификациями хроматина (Li *et al*., 2012). Рис. 8 показывает, что хроматин в контактирующих участках открыт – гистоны имеют маркеры модификации активной транскрипции – H3K4me3, H3K9ac (видны пики профиля), в то же время модификации репрессии транскрипции H3K27me3 не имеют пиков (равномерный шум).

Экспериментально полученные профили модификаций хроматина позволяют предска-

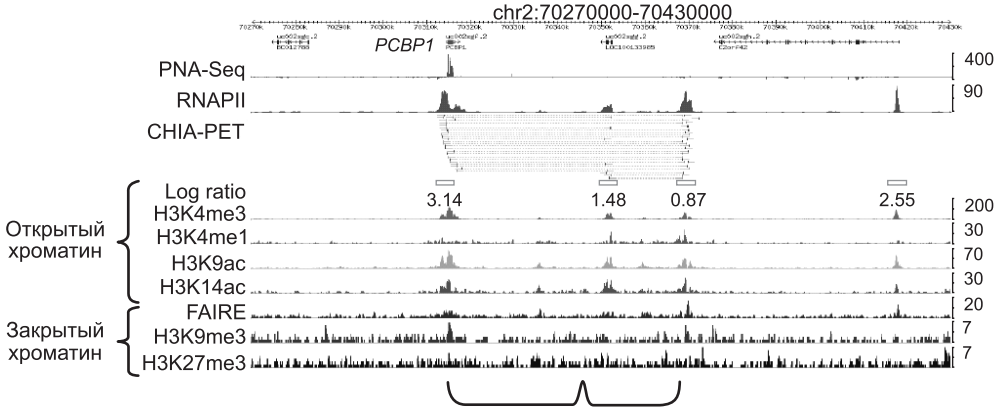

Рис. 8. ChIA-PET. Пример модификаций гистонов в участках хромосомных контактов (Li *et al*., 2012. P. 84–98).

зывать сайты связывания транскрипционных факторов в полногеномной шкале, что было показано детально для связывания рецептора эстрогенов ERα (Joseph *et al*., 2010). Более того, уже методом ChIA-PET показано, что ассоциации хромосомных контактов, опосредованных белком ERα, также связаны с открытым (активированным) состоянием хроматина, в частности модификациями H3K4me3, H3K4me1 (Fullwood *et al*., 2009) в линии клеток MCF-7.

ЗАКЛЮЧЕНИЕ

Технологии секвенирования, сопряженные с иммунопреципитацией хроматина (ChIP), позволяют исследовать механизмы регуляции транскрипции в масштабе генома эукариот (Collas, Dahl, 2008; Tucker *et al*., 2009). В данном обзоре представлены особенности моделирования такого рода полногеномных данных, связанные с анализом профилей ChIP-seq, выделением сайтов связывания и дальнейшим анализом их нуклеотидных последовательностей (Chen *et al*., 2008; Joseph *et al*., 2010; Li *et al*., 2012). Технология ChIP-seq позволяет исследовать профили связывания различных транскрипционных факторов и кофакторов, сопоставлять карты их взаимодействий с профилями модификаций хроматина и нуклеосомной упаковки (Chen *et al*., 2008). Расширение технологии секвенирования ДНК на пространственно-контактирующие участки хромосом (ChIA-PET) позволяет изучать пространственную организацию генома в ядре клетки также в связи с регуляцией транскрипции генов (Li *et al*., 2012). Рост объемов таких гетерогенных экспериментальных данных требует разработки новых биоинформационных решений (Laajala *et al*., 2009; Bailey *et al*., 2013).

БЛАГОДАРНОСТИ

Компьютерные расчеты проводились на оборудовании ССКЦ СО РАН. Работа выполнена при поддержке гранта РФФИ 12-04-00897-а, Программы Президиума РАН № 28 «Проблемы происхождения жизни и эволюция биосферы», Программы «Молекулярная и клеточная биология» – Интеграция РАН 6.6, Интеграционных проектов СО РАН № 39, 47, 136.

ЛИТЕРАТУРА

Bailey T., Krajewski P., Ladunga I. *et al.* Practical guidelines for the comprehensive analysis of chIP-seq data // PLoS Comput. Biol. 2013. V. 9. No. 11. e1003326.

Chen X., Xu H., Yuan P. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells // Cell. 2008. V. 133. No. 6. P. 1106–1117.

Chia N.-Y., Chan Y.-S., Feng B. *et al.* A genome-wide RNAi screen identifies PRDM14 as a regulator of POU5F1 and human embryonic stem cell identity // Nature. 2010. V. 468. No. 7321. P. 316–3120.

Collas P., Dahl J.A. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation // Front. Biosci. 2008. V. 13. P. 929–943.

Dekker J., Marti-Renom M.A., Mirny L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data // Nat. Rev. Genet. 2013. V. 14. No. 6. P. 390–403.

ENCODE Project Consortium / B.E. Bernstein, E. Birney, I. Dunham, E.D. Green, C. Gunter, M. Snyder. An integrated encyclopedia of DNA elements in the human genome // Nature. 2012. V. 489. No. 7414. P. 57–74.

Fullwood M.J., Liu M.H., Pan Y.F. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome // Nature. 2009. V. 462. No. 7269. P. 58–64.

Joseph R., Orlov Y.L., Huss M. Integrative model of genomic factors for determining binding site selection by estrogen receptor α // Mol. Syst. Biol. 2010. V. 6. P. 456.

Kaplan N., Moore I.K., Fondufe-Mittendorf Y. *et al.* The DNA-encoded nucleosome organization of an eukaryotic genome // Nature. 2009. V. 458. No. 7236. P. 362–366.

Kedes L., Campany G. The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE Competition // Nature Genet. 2011. V. 43. P. 1055–1058.

Kuznetsov V.A., Orlov Y.L., Wei C.L., Ruan Y. Computational analysis and modeling of genome-scale avidity distribution of transcription factor binding sites in chip-pet experiments // Genome Inform. 2007. V. 19. P. 83–94.

Laajala T.D., Raghav S., Tuomela S. *et al.* A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments // BMC Genomics. 2009. V. 10. P. 618.

Langmead B., Trapnell C., Pop M., Salzberg S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome // Genome Biol. 2009. V. 10. No. 3. R25.

Lee H., Schatz M.C. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score // Bioinformatics. 2012. V. 28. No. 16. P. 2097–2105.

Li R., Li Y., Kristiansen K., Wang J. SOAP: short oligonucleotide alignment program // Bioinformatics. 2008. V. 24. P. 713–714.

Li G., Ruan X., Auerbach R.K. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation // Cell. 2012. V. 148. No. 1/2. P. 84–98.

Liu X., Noll D.M., Lieb J.D., Clarke N.D. DIP-chip: rapid and accurate determination of DNA-binding specificity //

Genome Res. 2005. V. 15. No. 3. P. 421–427.
Tucker T., Marra M., Friedman J.M. Massively parallel sequencing: the next big thing in genetic medicine // Am. J. Hum. Genet. 2009. V. 85. No. 2. P. 142–154.
Wei C.L., Wu Q., Vega V.B. *et al.* A global map of p53 transcription-factor binding sites in the human genome // Cell. 2006. V. 124. No. 1. P. 207–219.
Zeller K.I., Zhao X., Lee C.W. *et al.* Global mapping of c-Myc binding sites and target gene networks in human B cells // Proc. Natl Acad. Sci. USA. 2006. V. 103. P. 17834–17839.
Zhang Y., Liu T., Meyer C.A. *et al.* Model-based analysis of ChIP-Seq (MACS) // Genome Biol. 2008. V. 9. No. 9. R137.

COMPUTER-ASSISTED STUDY OF THE REGULATION OF EUKARYOTIC GENE TRANSCRIPTION ON THE BASE OF DATA ON CHROMATIN SEQUENCING AND PRECIPITATION

Yu. L. Orlov

Institute of Cytology and Genetics, Siberian Branch
of the Russian Academy of Sciences, Novosibirsk, Russia;
e-mail: orlov@bionet.nsc.ru

The development of high-throughput sequencing technologies dramatically accelerated the accumulation of data on genome structure, distribution of gene regulatory regions in the genome and specific features of their interaction. Technologies associated with chromatin immunoprecipitation are reviewed: ChIP-PET, ChIP-seq, and ChIA-PET. Computer-assisted methods for the analysis of transcription factor binding sites (TFBSs) and regulatory region structures throughout the genome are described. The paper demonstrates approaches to the solution of tasks related to genomic data annotation and identification of TFBSs and regulatory regions.

Key words: sequencing, chromatin immunoprecipitation, ChIP-chip, ChIP-seq, ChIP-PET, ChIA-PET, transcription factor binding sites, gene expression regulation.