ГЕНЕТИКА, 2015, том 51, № 4, с. 409–429

УДК 575.113

РЕГУЛЯТОРНАЯ ГЕНОМИКА – ЭКСПЕРИМЕНТАЛЬНО-КОМПЬЮТЕРНЫЕ ПОДХОДЫ

© 2015 г. Е. В. Игнатьева[1, 2], О. А. Подколодная[1], Ю. Л. Орлов[1, 2], Г. В. Васильев[1], Н. А. Колчанов[1, 2]

[1]Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск 630090
[2]Новосибирский национальный исследовательский государственный университет, Новосибирск 630090
e-mail: eignat@bionet.nsc.ru

Поступила в редакцию 20.10.2014 г.

Настоящий обзор посвящен описанию экспериментально-компьютерных подходов к исследованию механизмов регуляции транскрипции и организации регуляторных районов генов эукариот, включая: а) изучение факторов, определяющих величину аффинности взаимодействия TATA-боксов к TBP (TATA-binding protein); б) исследование закономерностей распределения маркеров хроматина и их вклада в уровень экспрессии генов; в) изучение трехмерной структуры хроматина; г) анализ влияния нуклеотидных замен на экспрессию генов с использованием методов ChIP-seq и DNase-seq в рамках полногеномных экспериментов. Показано, что именно экспериментально-компьютерным подходам принадлежит ключевая роль в формировании современных представлений о механизмах регуляции транскрипции и структурно-функциональной организации регуляторных районов, контролирующих этот процесс.

DOI: 10.7868/S0016675815040062

В последнее десятилетие в геномных исследованиях произошла технологическая революция. Стремительно снижается стоимость расшифровки геномов на основе технологий секвенирования нового поколения (NGS), что привело к расшифровке десятков тысяч геномов различных видов эукариот и бактерий [1, 2]. NGS технологии широко применяются также для изучения изменчивости генома человека: на их основе реализован такой крупномасштабный проект как “1000 геномов”, в рамках которого выявлено ~20 млн новых однонуклеотидных замен, ~1 млн коротких делеций/инсерций, ~7000 крупных делеций [3].

На смену методам анализа транскриптома с помощью экспрессионных микрочипов приходят методы секвенирования полных транскриптомов клеток и тканей (RNA-Seq), дающие существенно более точные оценки уровня экспрессии транскриптов. Растет количество данных, полученных с использованием новых высокопроизводительных методов: CAGE, SAGE, RNA-PET и RNA-Seq для идентификации стартов транскрипции; ChIP-seq (хроматин-иммунопреципитация) для анализа модификаций гистонов и связывания хроматина с транскрипционными факторами; DNase-seq для выявления сайтов гиперчувствительности, соответствующих открытому хроматину, и др.

Эти и другие методы получения качественно новых знаний о транскрипционном уровне регуляции активности генов способствовали бурному развитию исследований в области регуляторной геномики и накоплению огромных объемов экспериментальных данных высокой сложности, понимание которых возможно только при тесной интеграции экспериментальных и биоинформатических подходов, новых информационных технологий, методов компьютерного анализа и математического моделирования.

Постоянно растущий интерес к изучению механизмов транскрипционного контроля экспрессии генов объясняется тем, что транскрипция является ключевым событием, инициирующим сложный многостадийный процесс экспрессии генов эукариот, включающий помимо транскрипции такие этапы, как процессинг РНК, трансляция, посттрансляционная модификация белка и т.д.

Настоящий обзор посвящен описанию экспериментально-компьютерных подходов к исследованию механизмов регуляции транскрипции, а также организации регуляторных районов генов эукариот. Будут рассмотрены результаты экспериментально-компьютерных работ в различных областях регуляторной геномики, включая: а) исследование характеристик TATA-боксов, определяющих величину аффинности TBP (TATA-binding protein), инициирующего сборку преинициационного транскрипционного комплекса (**ПИК**); б) изучение закономерностей распределения маркеров хроматина и их вклада в уровень экспрессии генов; в) экспериментальные и теорети-

ческие подходы к исследованию трехмерной структуры хроматина; г) анализ влияния нуклеотидных замен на экспрессию генов с использованием методов ChIP-seq и DNase-seq в рамках полногеномных экспериментов. Как будет показано в обзоре, именно экспериментально-компьютерным подходам принадлежит ключевая роль в формировании новых представлений о механизмах регуляции транскрипции и структурно-функциональной организации регуляторных районов, контролирующих этот процесс.

МЕТОДЫ ИССЛЕДОВАНИЯ РЕГУЛЯТОРНЫХ РАЙОНОВ, КОНТРОЛИРУЮЩИХ ТРАНСКРИПЦИЮ, ОСНОВАННЫЕ НА ВЫСОКОПРОИЗВОДИТЕЛЬНОМ СЕКВЕНИРОВАНИИ

Все современные подходы, применяемые для решения задач регуляторной геномики, опираются на технологии высокопроизводительного секвенирования ДНК второго либо третьего поколения. Радикально снизив стоимость ресеквенирования геномов, эти технологии дали возможность перевода в полногеномный формат методов изучения регуляторных геномных последовательностей, контролирующих процессы транскрипции генов.

Метилирование ДНК

Основным инструментом исследования статуса метилирования ДНК, играющего важную роль в регуляции транскрипции и эпигенетическом программировании геномов, является бисульфитное секвенирование, заключающееся в обработке ДНК бисульфитом натрия с последующим секвенированием геномного материала и компьютерным выравниванием полученных последовательностей на референсный геном [4]. Метод позволяет различать метилированные и неметилированные цитозины, зачастую давая заметный разброс в оценке экспериментальных данных [5]. Поэтому находят применение и другие методы, основанные на использовании рестрикционных ферментов, чувствительных к метилированным цитозинам [6], либо антител, специфичных к метилированной ДНК [7]. Представляет интерес комбинирование методов бисульфитного секвенирования и иммунопреципитации хроматина (ChIP-seq) [8].

Иммунопреципитация хроматина

Иммунопреципитация хроматина (ChIP-seq – Chromatin ImmunoPrecipitation) с последующим секвенированием – распространенный метод выявления ДНК-белковых взаимодействий в хроматине, основанный на обработке клеток формальдегидом, приводящей к образованию ковалентных сшивок между ДНК и белками [9]. Обработанный ядерный хроматин дробится на фрагменты длиной 250–500 пн (рис. 1). Затем с помощью антител, специфичных к целевым белкам, выделяются сшитые ДНК-белковые комплексы; далее ДНК выделяется из комплексов и секвенируется.

Существенной частью этого подхода является компьютерный анализ результатов секвенирования (рис. 1). Компьютерное картирование секвенированных последовательностей ДНК на геном представляет собой достаточно объемную задачу биоинформатики, требующую соответствующих вычислительных ресурсов [10], использования различных форматов платформ секвенирования, в том числе форматов цветовой кодировки SOLiD [11].

Однозначность картирования представляет отдельную проблему анализа данных. Пример затруднений – картирование фрагментов ДНК в генах, для которых известны псевдогены. Существует понятие “картируемости” (mappablity) как свойства нуклеотидных последовательностей хромосом в геноме, определяемое однозначностью расположения коротких последовательностей заданной длины [12]. Для каждой длины последовательности ДНК существует своя “уникома” – например, для фрагментов размером 50 нуклеотидов некартируемых участков гораздо меньше, чем для фрагментов длиной 25 нуклеотидов. Существуют готовые разметки – профили “уникальности” для нескольких референсных геномов, в частности геномов человека и мыши [12].

Используя координаты секвенированных фрагментов ДНК на хромосомах референсного генома, строится численный профиль ChIP-seq, определяются его пики. Высота пика измеряется количеством выравниваний секвенированных фрагментов ДНК в соответствующей точке генома. Качество сигнала связывания белка с ДНК в профиле ChIP-seq оценивается через отношение числа специфичных фрагментов ДНК (связанных с белком) в рассматриваемой точке генома к числу неспецифичных фрагментов, полученных в контрольном эксперименте ChIP-seq с использованием в качестве антител иммуноглобулина IgG или GFP, не имеющих специфического связывания с ДНК [13].

Выделение пиков в профиле ChIP-seq требует специализированных компьютерных программ, ориентированных на конкретную задачу, в зависимости от а) технологий секвенирования (коррекция на специфические ошибки), б) размера и особенностей эукариотического генома (наличие повторенных последовательностей, детали аннотации). Секвенирование фрагментов ДНК может выполняться не только с одного, но и с двух концов (с использованием технологии PET – Paired

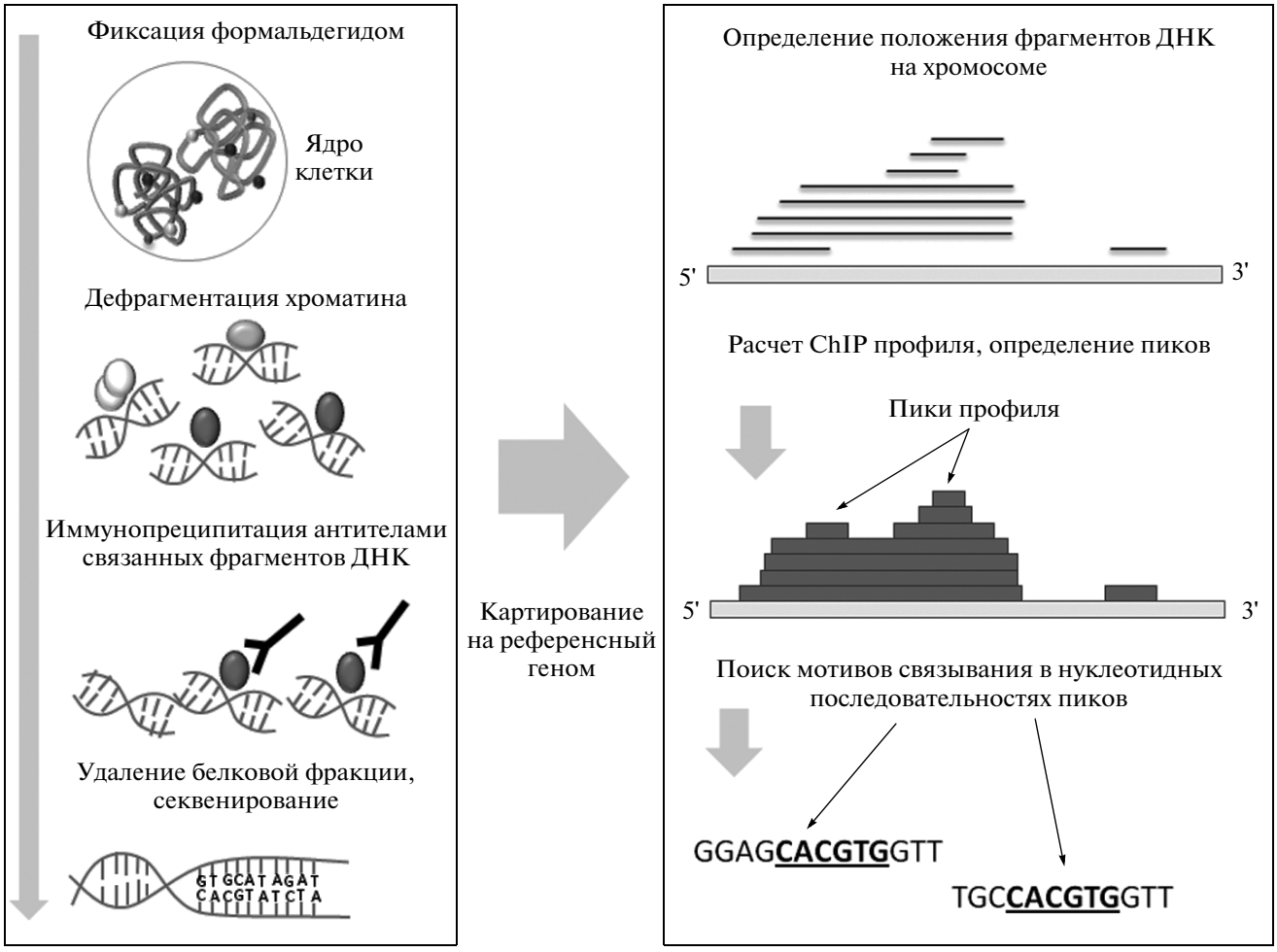


Рис. 1. Схема эксперимента ChIP-seq и анализа геномных профилей.

End Tags), позволяя более точно картировать сайты связывания транскрипционных факторов (**ТФ**) [14, 15], что требует специальных программ для одновременного картирования пар прочтений [16].

Для оценки качества сигнала связывания в профиле ChIP-seq и выделения набора пиков с использованием контрольных профилей разработан ряд компьютерных программ: GLITR, MACS, HPeak, PeakFinder, GLITR, QuEST, CisGenome, USeq и PICS [13, 17].

Программа MACS [17] 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.

Для картирования фрагментов ДНК на геном используются программы MAQ (Mapping and Assembly with Quality), SOAP (Short Oligonucleotide Alignment Packet) [18], ELAND. Программа ELAND ориентирована на стандарт данных, полученных на аппаратуре Illumina (http://www.illumina.com/systems.ilmn).

Распространены программы Bowtie, SeqMap, RMAP, ZOOM [13, 19]. Для анализа участков связывания РНК-полимеразы и выделения промоторных районов разработан метод GRO-Seq (Global Run-On sequencing), использующий те же компьютерные технологии анализа профилей связывания, что и ChIP-seq [20, 21].

Завершающим этапом анализа пиков профиля ChIP-seq является идентификация сайтов связывания ТФ в пределах этих пиков. Решается задача выявления точного положения сайтов связывания в пределах пиков с использованием баз данных нуклеотидных мотивов, соответствующих ТФ, описываемым консенсусами, весовыми матрицами и др. [22], а также с использованием экспериментально подтвержденных моделей сайтов связывания транскрипционных факторов [23].

Изучение районов активного хроматина

Метод DNase-seq основан на обработке клеток ДНКазой I, преимущественно разрушающей ДНК транскрипционно-активных районов генома, слабо экранированных белками хроматина. Последующее выделение пула образующихся коротких фрагментов ДНК (~20 пн либо ~50 пн) с их секвенированием и компьютерное выравнивание на геномную последовательность позволяют локализовать транскрипционно-активные районы с расположенными в них регуляторными элементами [24].

Исследование транскриптома

Первым методом изучения транскриптома стал SAGE (serial analysis of gene expression). На выделенной поли-А РНК синтезируется кДНК, которая в ряде последовательных ферментативных обработок превращается в короткие фрагменты со специфичными линкерами по концам, центральные районы которых (таги) соответствуют коротким участкам транскриптов (вставкам размером в 20 пн в варианте LongSAGE, либо 26 пн в варианте SuperSAGE) [25]. Далее эти короткие фрагменты лигируются в длинные конкатемеры и секвенируются. Компьютерный анализ позволяет на основе распределения тагов по геному оценивать спектр экспрессирующихся генов, а также и уровни их экспрессии (по количеству секвенированных тагов). Несмотря на появление более совершенных подходов, метод SAGE до сих пор находит применение для исследования транскриптома [26].

Наиболее часто используемым является RNA-Seq – метод секвенирования всего пула клеточных РНК, расщепленных на короткие фрагменты. Поскольку рибосомальная РНК составляет, как правило, более 90% всей клеточной РНК, она обычно удаляется с помощью гибридизации. Следует различать профилирование экспрессии генов, при котором используется только поли-А РНК, и собственно RNA-Seq, при котором исследуется весь пул клеточных РНК, включая некодирующие. С помощью RNA-Seq были обнаружены огромное разнообразие вариантов альтернативной транскрипции длинных некодирующих РНК [27] и новый класс кольцевых регуляторных РНК [28] у эукариот, а также сложная организация бактериального транскриптома (включающая альтернативную транскрипцию генов, наличие антисмысловых и микроРНК и даже вырезаемых интронов) [29].

Идентификация сайтов начала транскрипции

Для массового обнаружения стартов транскрипции разработан метод 5' RACE, основанный на быстрой амплификации 5'-концов кДНК. Для синтеза кДНК используется праймер к известной части гена, далее к 3'-концу кДНК достраивается гомополимерный участок, служащий в качестве второго праймера в ПЦР. Полученные фрагменты секвенируются и анализируются путем компьютерного выравнивания с референсным геномом. Этот метод используется в основном при изучении транскриптомов прокариот.

Для изучения стартов транскрипции у эукариот широко применяется метод CAGE – кэп-анализ экспрессии генов. Короткие участки длиной 27 нуклеотидов (начиная от 5'-конца кэпированной РНК) используются для синтеза кДНК, амплифицируются и секвенируются на высокопроизводительных секвенаторах с компьютерным выравниванием на известную референсную геномную последовательность. Метод позволяет получать данные не только о положении стартов транскрипции, но и об относительных уровнях экспрессии, хорошо дополняя данные RNA-Seq [30].

Для уменьшения искажений, вносимых ПЦР, разработаны варианты CAGE без амплификации ДНК в ходе пробоподготовки и ориентированные на использование секвенаторов третьего поколения (Helicos) [31] и второго поколения (Illumina) [32].

ПОЛНОГЕНОМНЫЕ ПРОЕКТЫ, НАПРАВЛЕННЫЕ НА ИССЛЕДОВАНИЕ РЕГУЛЯТОРНЫХ РАЙОНОВ, КОНТРОЛИРУЮЩИХ ТРАНСКРИПЦИЮ

Широкомасштабные проекты геномных исследований делают возможной детальную функциональную аннотацию регуляторных геномных последовательностей по экспериментальным данным, полученным в результате массового параллельного секвенирования ДНК [33, 34]. Кроме определения положения и структуры белок-кодирующих генов, полногеномная аннотация включает описание некодирующих РНК, выделение регуляторных районов генов, исследование хромосомных аномалий и однонуклеотидных замен, определение функции белков, предсказание их вторичной и пространственной структуры [35, 36].

Энциклопедия элементов ДНК (англ. The Encyclopedia of DNA Elements, ENCODE) создана в рамках Международного консорциума, организованного с целью детального анализа функций элементов генома человека. В 2012 г. первые результаты проекта были опубликованы в виде 30 взаимосвязанных публикаций на сайтах журналов “Nature”, “Genome Biology” и “Genome Research” [36]. Показано, что большая часть генома человека, до 80%, имеет биологические функции (биохимическую активность). До этого господствовало представление о том, что большая часть ДНК является “избыточной”.

По аналогии с проектом ENCODE осуществляется проект modENCODE [37] – картирование функциональных элементов генома основных модельных объектов – *D. melanogaster* и *C. elegans* (modENCODE – от англ. Model Organism ENCyclopedia Of DNA Elements). Достоинством данного проекта является возможность проведения на модельных организмах таких экспериментов, которые трудно или невозможно осуществить на человеке [38, 39]. В 2010 г. консорциум modENCODE представил ряд статей по аннотации и анализу распределения функциональных элементов в геноме *D. melanogaster* и *C. elegans*, эти исследования продолжаются с использованием дополнительных данных [40, 41].

Широкомасштабное исследование транскриптомов началось с проекта FANTOM [42] с помощью секвенирования методами CAGE, SAGE и MPSS [30]. Анализ полноразмерных кДНК в геноме мыши позволил выявить более 180 тыс. различных вариантов транскриптов, отличающихся использованием альтернативных промоторов, сплайсингом, участками полиаденилирования [43]. При исследовании транскриптома млекопитающих обнаружено явление цис-антисенс транскрипции, когда транскрипция с цепи ДНК идет в противоположных направлениях [44, 45]. На современном этапе выполнения проекта (FANTOM5) методом CAGE с использованием одномолекулярных секвенаторов третьего поколения были получены данные о позициях стартов транскрипции и уровнях экспрессии генов в 975 типах (нормальных и раковых) клеток человека и 399 типах клеток мыши. Анализ этих данных показал, что большинство промоторов млекопитающих содержат множественные близкорасположенные старты транскрипции, каждому из которых соответствует индивидуальный профиль активности в различных типах клеток [46]. Кроме того, в рамках проекта FANTOM5 была охарактеризована активность более 43 тыс. энхансеров человека в 432 типах клеток из первичных культур, а также в широком круге тканей (135 типов) и 241 линии клеток [47].

ИССЛЕДОВАНИЕ ТРЕХМЕРНОЙ СТРУКТУРЫ ХРОМАТИНА

Трехмерная структура генома активно исследуется при помощи различных методов, основанных на геномном секвенировании: 3C, Hi-C, ChIA-PET и др. (рис. 2). Метод Hi-C основан на изучении хромосомных контактов по технологиям 3C (Chromosome Conformation Capture) и массового параллельного секвенирования. На первом этапе, как и в методе ChIP-seq, проводится обработка формальдегидом с образованием белок-нуклеиновых сшивок и извлечение комплексов с помощью антител против определенного ядерного белка (рис. 2). Фрагментация проводится с помощью рестриктаз, создающих достаточно длинные фрагменты, которые в условиях сильного разбавления лигируются сами на себя до расшивки комплексов. В таких условиях лигируются только концы молекул ДНК, сближенные в пространстве. Далее нужные молекулы ДНК выделяются (обычно с помощью мечения биотином) и подвергаются секвенированию. Затем осуществляется компьютерное выравнивание секвенированных фрагментов ДНК, каждый из которых образован парой удаленных по геному, но сближенных в пространстве районов, оказавшихся в результате описанной выше процедуры в составе одного и того же секвенированного фрагмента ДНК. Такие участки идентифицируются как места хромосомных контактов [48, 49].

В методе Hi-C [50] выделение ДНК-белковых комплексов проводится без использования специфичных антител, что требует в последующем гораздо больших объемов секвенирования, но позволяет получить более общее представление о пространственном расположении отдельных хромосом относительно друг друга в интерфазном ядре.

Исследование хромосомных контактов ставит задачу их статистического анализа [48, 51], требующую обработки больших объемов экспериментальных данных, превышающих стандартные объемы данных ChIP-seq. При обработке данных Hi-C выделяются пространственные домены на хромосоме [52]: такая информация представлена в Интернет-доступных базах данных (например, 3DGD и Mouse Encode Project at Ren Lab).

При помощи метода Hi-C было подтверждено наличие в ядре хромосомных территорий [53]. Отмечается, что межхромосомные контакты очень динамичны, т.е. в различных клетках одной и той же клеточной популяции распределение контактов может значительно различаться [54, 55]. Метод ChIA-PET (Chromatin Immunoprecipitation Analysis – Paired End Tags) [48, 56] использует иммунопреципитацию хроматина, позволяя определять сближенные участки хромосом, контакты которых опосредованы определенным белком – транскрипционным фактором или белковым комплексом (например, ERα или CTCF) [56].

Компьютерный анализ распределения хромосомных контактов относительно генов и участков генома, ассоциированных с модификациями хроматина (модификации гистона H3), выявил связь контактирующих участков с открытым состоянием хроматина [57].

Предложена классификация моделей промоторных, энхансерных и мультигенных контактов, опосредованных комплексом РНК-полимеразы II [48]. В ней выделены классы: базальный

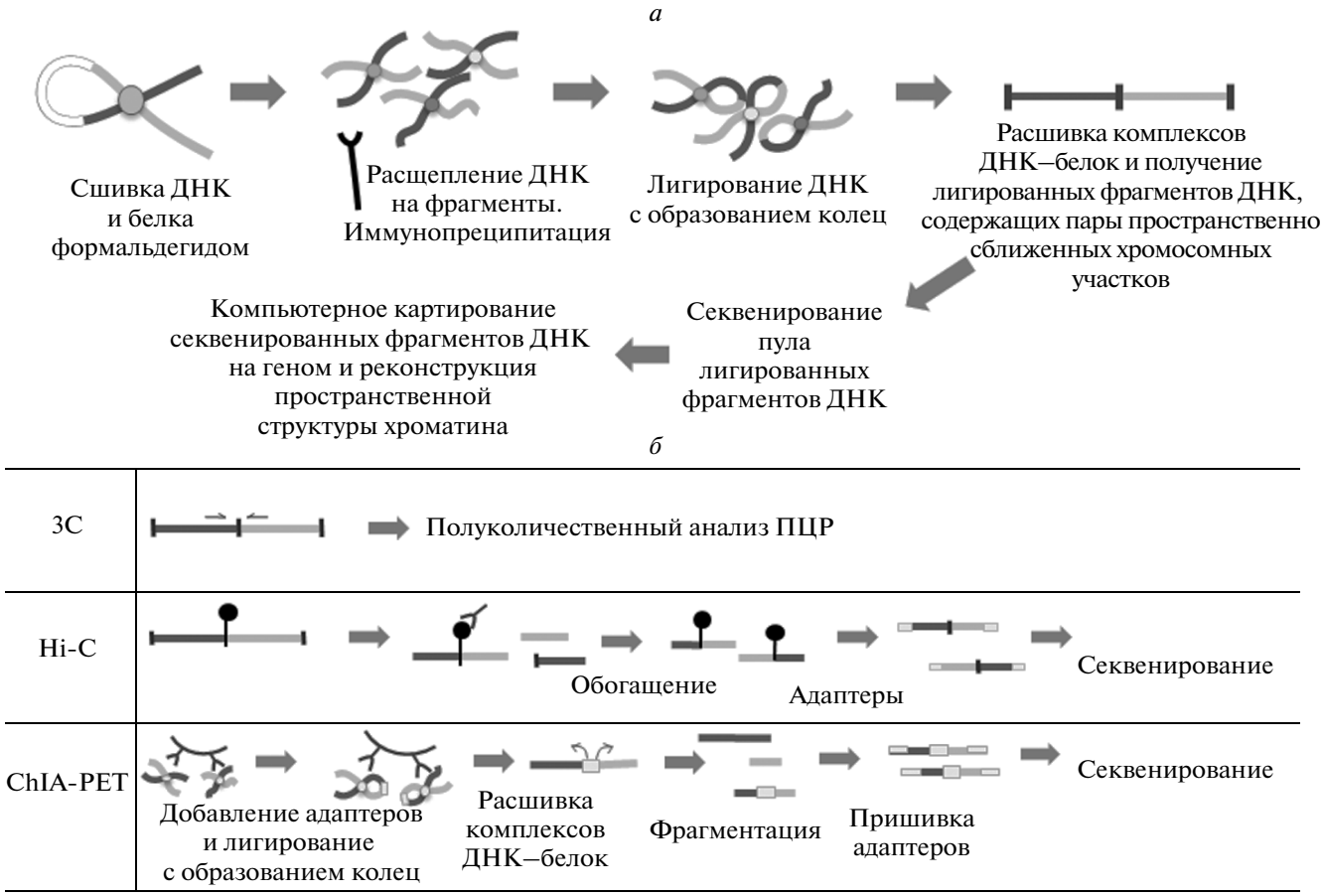


Рис. 2. Методы анализа хромосомных контактов с помощью секвенирования: 3C, Hi-C и ChIA-PET. *а* – общие этапы анализа; *б* – этапы, специфичные для каждой отдельной методики. Представлено по [49] с модификациями.

промотор, промотор-энхансер и мультигенный локус. Модель базального промотора включает только локальные петли ДНК в промоторе, без удаленных взаимодействий. В модели одиночного гена учитываются только петли в районе гена – между энхансером и промотором, возможно между 5'- и 3'-районами гена, но без других белок-кодирующих генов. Мультигенная модель включает сразу несколько генов, расположенных рядом друг с другом на хромосоме и контактирующих промоторными районами. Для мультигенной модели введен термин “хромоперон” (chromoperon) или хромосомный оперон [48].

Хотя потенциальные энхансеры могут быть определены экспериментально [58], остается нерешенной проблема сопоставления энхансеров и их генов-мишеней, находящихся на удалении сотен тысяч нуклеотидов [59]. Более того, многие дальние энхансеры могут быть вложены в интронные районы других дистально расположенных генов [60], что затрудняет соотнесение энхансеров к их генам-мишеням. С помощью ChIA-PET было идентифицировано приблизительно 1000 ультра-дальнодействующих (расположенных далее 500000 пн) энхансер-промоторных взаимодействий, которые также специфичны для клеточных линий человека [48].

БАЗЫ ДАННЫХ, МАССИВЫ ДАННЫХ, БРАУЗЕРЫ

Базы данных, содержащие информацию по регуляции транскрипции, начали создаваться с 90-х годов XX в. Первыми широко известными и признанными мировым сообществом базами были ooTFD, EPD, TRANSFAC, TRRD [61–64]. Эти базы наполнялись путем ручной аннотации научных публикаций и содержали сведения о регуляторных районах генов, ТФ и их сайтах связывания на ДНК. Позднее, с развитием массовых методов анализа регуляторных районов генов, была разработана японская база DBTSS, включающая данные о позициях стартов транскрипции генов и характеристики регуляторных районов различных организмов, включая человека и мышь. Массивы данных для базы DBTSS были собраны в ходе выполнения нескольких широкомасштабных геномных экспериментов с использованием разных

методов анализа (TSS-seq, RNA-Seq, ChIP-seq, BS-seq) и разных типов клеток, включая раковые [65]. Базы TRED и MPromDb были созданы на основе интеграции данных из других информационных источников, а также компьютерного предсказания регуляторных элементов [66, 67]. Согласно сведениям, представленным на сайте журнала Nucleic Acids Research (http://www.oxfordjournals.org/nar/database/c/), в 2014 г. количество баз данных по регуляции транскрипции достигло восьмидесяти [68]. К этой категории отнесены также информационные ресурсы по регуляторным белкам (транскрипционным факторам и белкам с корегуляторной активностью). Наиболее значимыми являются AnimalTFDB, TFClass, CREMOFAC [69–71]. К числу ресурсов по тематике регуляции транскрипции причислены также базы, включающие матрицы сайтов связывания ТФ – JASPAR и HOCOMOCO [72, 73].

Огромные массивы информации доступны через Интернет-сайты проектов ENCODE (http://genome.ucsc.edu/ENCODE/), FANTOM (http://fantom.gsc.riken.jp/), modENCODE (http://www.modencode.org/) и ряда других. Разработчики этих проектов предоставляют возможность доступа к информации через специальные браузеры, сервисы (ftp-сайты либо таблицы), либо создают каталоги статей, опубликованных по проектам и включающих огромные массивы данных в форме приложений.

РЕГУЛЯТОРНЫЕ РАЙОНЫ ГЕНОВ ЭУКАРИОТ

Новые экспериментальные технологии в сочетании с компьютерными подходами существенно расширили наши знания о регуляторных районах генов эукариот, выполняющих разные функции (промоторах, энхансерах, инсуляторах).

Промоторы

Промотор можно определить как область ДНК, непосредственно окружающую сайт инициации транскрипции (TSS), где происходит сборка преинициационного комплекса, а также включающую близлежащие последовательности, которые интегрируют сигналы, регулирующие работу гена [74]. Для большинства генов млекопитающих характерно наличие множественных промоторов, альтернативное использование которых способствует формированию разнообразия и сложности транскриптома и протеома [75]. Промоторы могут характеризоваться особенностями структурной организации и набора цис-регуляторных элементов. Исходя из структурной организации, можно выделить два основных типа промоторов. К первому типу относятся промоторы с фокусированным стартом транскрипции, т.е. такие, где транскрипция начинается с одного или реже с нескольких близкорасположенных нуклеотидов. Ко второму типу относятся промоторы с множественными стартами транскрипции в пределах сегментов длиной около 100 пн, которые, как правило, располагаются в CpG островках. У позвоночных первая группа составляет менее 30%, вторая – более 70% генов [76]. Вопрос о механизме предпочтения того или иного старта транскрипции в промоторах второго типа остается открытым. Недавние результаты, полученные в рамках выполнения проекта FANTOM5, позволяют предполагать, что ключевым индикатором предпочтения определенного старта транскрипции в промоторах с множественными стартами транскрипции является позиционированная нуклеосома [46].

Область 35–40 пар нуклеотидов выше или ниже старта транскрипции называется базальным промотором (core promoter), содержащим короткие мотивы ДНК, которые взаимодействуют с транскрипционной машиной, в том числе с TFIID, РНК-полимеразой II. Эти мотивы выполняют задачу правильного позиционирования ее в районе старта транскрипции.

В качестве примера некоторых из этих мотивов можно назвать: TATA-бокс (консенсусная последовательность TATAWAAR; расстояние от старта транскрипции – 30 пн), Inr (YYANWYY (*H. sapiens*), TCAKTY (*D. melanogaster*); –2/+4), DPE (RGWYVT; +28/+33), MTE (CSARCSSAACGS; +18/+29), BRE^{u} (SSRCGCC; выше TATA-бокса), BRE^{d} (RTDKKK; –23/–17), XCPE1 (DSGYGGRASM (*H. sapiens*); –8/+2) [76, 77]. Данные мотивы не являются универсальными, общими для всех промоторов; невозможно выделить какой-либо элемент, абсолютно необходимый для функционирования базального промотора, так как промотор может содержать только один регуляторный элемент, либо определенную комбинацию элементов, что, вероятно, сказывается как на активности, так и на характере тонкой регуляции этих промоторов. Значительная часть выявленных промоторов с дисперсным стартом не содержит ни одного из описанных мотивов [78, 79].

Энхансеры

Энхансер – это последовательность элементов ДНК, которая стимулирует активность связанного с ним промотора независимо от ориентации [80]. Как правило, энхансеры представляют собой районы ДНК в несколько сотен пар оснований, содержащие сайты связывания широкого спектра ТФ (активаторов, репрессоров, модификаторов хроматина), сочетание которых обеспечивает разнообразие возможностей регуляции гена в тканеспецифичной манере в соответствии со стадией развития организма и необходимостью отве-

та на внешние стимулы. Энхансеры могут локализоваться в 5'- и 3'-областях генов и в их интронах на расстояниях от нескольких сотен до нескольких сотен тысяч пар оснований и даже на другой хромосоме, чем активируемый ими промотор [80, 81].

В настоящее время выявлено несколько типов особых энхансеров: суперэнхансеры, протяженные энхансеры, теневые энхансеры, сплит энхансеры. Типичным представителем суперэнхансеров являются локус-контролирующие районы, выявленные впервые в β-глобиновом локусе. Понятие "теневые энхансеры" появилось в результате полногеномного профилирования сайтов связывания ТФ, когда было выявлено наличие функционально активных энхансеров, расположенных на большом расстоянии от регулируемого гена и дублирующих действие основных энхансеров. Теневые энхансеры, как правило, регулируют транскрипцию генов развития и необходимы в неоптимальных условиях развития организма [82]. Сплит энхансеры работают в паре, но в отличие от теневых оба абсолютно необходимы для транскрипции гена, неправильная регуляция которого может быть летальна для организма [57].

Общепринятой моделью действия энхансеров является то, что энхансер за счет выпетливания ДНК образует контакт с промотором и далее способствует стабилизации связывания РНК-полимеразы или освобождению остановившейся полимеразы. Энхансеры маркируются в геноме специфическими наборами модификаций гистонов, такими как H3K4me и H3K27ac [36]. В то же время нужно отметить, что механизм действия энхансеров до сих пор слабо изучен. Недавние исследования по широкомасштабному геномному профилированию показали, что энхансеры часто, если не всегда, участвуют в транскрипции РНК, получившей название эРНК [47, 83]. В рамках проекта FANTOM5 на панели образцов первичных клеток из тканей человека был создан атлас активных энхансеров, транскрибируемых *in vivo*. Энхансеры были охарактеризованы как CpG обедненные промоторы, которые двунаправленно продуцируют относительно короткие (до 350 пн) несплайсируемые РНК, чувствительные к экзосомам. Генерация эРНК строго связана с активностью энхансера [47]. В то же время функциональная значимость эРНК остается неопределенной.

Инсуляторы

Инсуляторные элементы впервые описаны у дрозофил, и были охарактеризованы два их свойства – способность препятствовать распространению гетерохроматина и блокировать энхансерную активность в случае, когда такой элемент расположен между промотором и энхансером. С инсуляторными элементами дрозофил связывается целая группа белков, в частности Su(Hw), Zw5, CTCF, GAF, Mod(mdg4), BEAF-32 [84]. Из белков такого рода у млекопитающих наиболее изучен белок CTCF, который в комплексе с различными партнерами способен выполнять разнообразные функции [85]. Другим белком с инсуляторной функцией у млекопитающих является TFIIIC [86]. Для защиты от распространения гетерохроматина в эухроматиновые области инсуляторные элементы могут привлекать на соседние нуклеосомы факторы "активирующей" модификации гистонов (ацетилирование гистонов H3 и H4, метилирование H3K4), блокирующие распространение "сайленсинговых" гистоновых меток (метилирование H3K9 и K27) [87]. CTCF в комплексе с когезином вовлечен в образование петель хроматина, что может быть одним из механизмов, обеспечивающих блокирование инсулятором эффекта энхансера на промотор гена, в том случае когда эти два элемента оказываются разделенными в разные домены хроматина. Функция CTCF в составе инсулятора у позвоночных может быть как конститутивной, так и регулируемой. Конститутивный вариант наблюдается, например, в β-глобиновом локусе, где инсулятор, локализованный в конститутивном 5'HS5, блокирует действие энхансера на вышележащий ген и предотвращает распространение гетерохроматина в сторону β-глобиновго локуса [88]. Регулируемое блокирование энхансера инсулятором может определяться регулированием связывания CTCF с ДНК [89, 90]. В настоящее время появляется все больше данных в поддержку мнения, что инсуляторы являются регуляторными последовательностями, которые модулируют различные ядерные процессы, способствуя взаимодействиям между удаленными сайтами в геноме [91].

ЭКСПЕРИМЕНТАЛЬНО-КОМПЬЮТЕРНЫЕ ПОДХОДЫ К ИССЛЕДОВАНИЮ ТАТА-БОКСА И ЭНЕРГИИ СВЯЗЫВАНИЯ С TBP

Рассмотрим экспериментально-компьютерные подходы к исследованию TATA-бокса и энергии связывания с TBP. TATA-бокс, расположенный примерно 30 пн выше старта транскрипции, является одним из наиболее интенсивно изучаемых элементов базального промотора. Канонический TATA-бокс выявлен примерно в 10–16% промоторов у дрожжей [92] и 10% промоторов в геноме человека [93].

Механизм взаимодействия TATA-бокса с TBP-белком основан на одномерной диффузии TBP-белка вдоль ДНК (с интервалом скольжения ~1.5 тыс. пн) и осуществляется в несколько этапов. Первоначально белок TBP неспецифическим образом взаимодействует с любым участком на ДНК и в силу слабой аффинности к ДНК на-

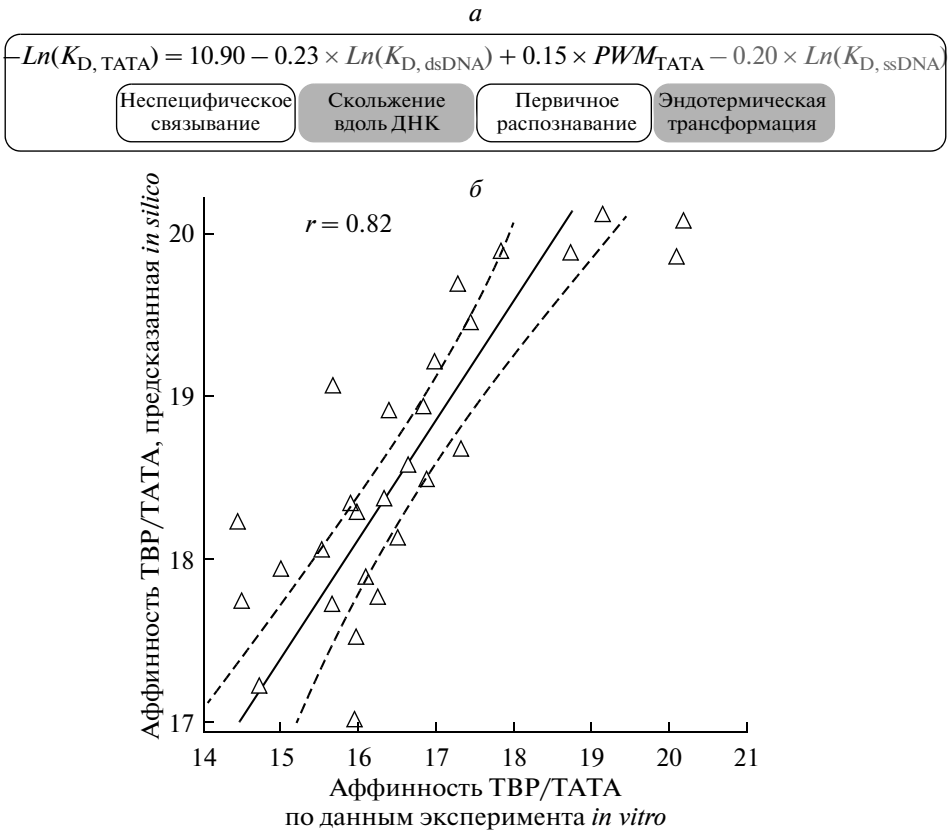


Рис. 3. Математическая модель для вычисления величины аффинности TBP белка человека к TATA-боксам ($-Ln(K_D)$): *а* – на основе их нуклеотидных последовательностей [96] и *б* – ее экспериментальное подтверждение по данным [98]. Сплошной линией обозначена линия регрессии; пунктирными линиями ограничен 95%-ный доверительный интервал.

чинает перемещаться вдоль нити ДНК [92, 94]. При контакте с TATA-боксом происходит первичное распознавание. На завершающем этапе происходит эндотермическая трансформация комплекса TBP/TATA, что приводит к изгибу ДНК на ~80° (у дрожжей) и ~105° у человека и обеспечивает стабилизацию комплекса TBP/TATA [95].

В работах М.П. Пономаренко с соавт. [96, 97] была разработана математическая модель взаимодействия и предложена формула расчета аффинности TBP/TATA (рис. 3,*а*). Для верификации модели аффинность TBP/TATA была исследована экспериментально *in vitro* методом задержки комплекса ДНК/белок в геле (EMSA) с использованием рекомбинантного человеческого TBP и возрастающих (вплоть до насыщения) количеств TATA-содержащих олигонуклеотидов. В этом эксперименте исследовали сродство TBP человека к синтетическим двуцепочечным олигонуклеотидам длиной 26 пн, последовательности которых соответствовали участкам, включающим TATA-боксы генов человека и их аллельные варианты. Был выявлен высокий уровень корреляции (r = 0.822) между величинами предсказанной и наблюдаемой аффинности (рис. 3,*б*), что подтверждает истинность предложенной математической модели [98]. С помощью данной математической модели была рассчитана аффинность природных вариантов TATA-боксов, ассоциированных с заболеваниями человека (альфа-, бета- и дельта-талассемией, инфарктом миокарда, тромбозом, нарушением иммунного ответа, амиотрофическим боковым склерозом, раком легких и гемофилией В Лейдена). Было предсказано, а впоследствии подтверждено экспериментально, что в большинстве случаев мутации вызывают снижение аффинности в 2–4 раза [98]. Более детальное экспериментальное исследование кинетических характеристик формирования комплекса TBP/TATA для нормальных и мутантных вариантов TATA-боксов показало, что мутантные формы TATA-боксов имеют сниженную (в 8–36 раз) константу скорости ассоциации (k_a) по сравнению с TATA-боксами здоровых индивидов. В таком случае вероятность образования комплекса TBP/TATA существенно снижается [99].

Таким образом, результаты, полученные в этом цикле работ, подтверждают высокую предсказательную силу формулы для расчета аффинности TBP/TATA и показывают, что эта формула

может быть использована для анализа последовательностей ТАТА-боксов и выявления функционально значимых однонуклеотидных замен.

ЭКСПЕРИМЕНТАЛЬНО-КОМПЬЮТЕРНЫЕ ПОДХОДЫ К ИССЛЕДОВАНИЮ СТРУКТУРЫ ХРОМАТИНА

Код модификаций хроматина

Дифференциальная экспрессия генов многоклеточного организма обеспечивается за счет регуляторных кодов транскрипции, объединяющих, помимо кода, записанного в форме сигналов на ДНК (наборы сайтов связывания транскрипционных факторов в регуляторных районах генов), также и ряд других (код позиционирования нуклеосом, код модификаций хроматина, включающий гистоновый код и код модификаций ДНК, код наднуклеосомной укладки хроматина) [77]. Регуляторные коды, взаимодействуя друг с другом [100, 101], обеспечивают интегральную регуляцию транскрипционной активности генов.

Для активной транскрипции необходимо, чтобы ДНК в составе хроматина была максимально доступна для регуляторных белков и РНК-полимеразы (открытый хроматин). Наиболее распространенными модификациями открытого хроматина является ацетилирование гистонов H3 и H4, ди- и триметилирование H3K4 и триметилирование H3K36 [100, 102].

Закрытый хроматин (гетерохроматин) может включать в себя гены, репрессированные на определенных этапах индивидуального развития либо в определенных типах клеток (факультативный гетерохроматин) [100].

Код модификаций хроматина имеет сложную структуру. Во-первых, нуклеосомы могут включать различные варианты гистоновых белков (например, H2A.Z вместо H2A или H3.3 вместо H3), различающиеся по пространственной структуре либо по концевым районам полипептидых цепей и выполняющие специализированные функции [102]. Во-вторых, можно наблюдать большое разнообразие возможных модификаций гистоновых белков: ацетилирование, метилирование, фосфорилирование, АДФ-рибозилирование, деиминирование, убиквитинирование, сумоилирование, присоединение N-ацетилглюкозамина, удаление концевых участков гистоновых белков (Histone tail clipping), изомеризация пролина [100]. В-третьих, спектр вариантов модификаций гистонов исключительно велик: 1) модификациям могут подвергаться аминокислотные остатки как на N-, так и на C-концевых участках гистонов; 2) каждый гистон может иметь модификации по нескольким позициям; 3) боковая цепь остатка лизина может метилироваться несколько раз (моно-, ди-, триметилирование). В-четвертых, модификациям подвергаются не только гистоновые белки, но и ДНК: возможны такие модификации цитозина, как метилирование, 5'-гидроксиметилирование, карбоксилирование и формилирование (присоединение формильной группы) [101, 103].

О сложности кода модификаций хроматина свидетельствует также тот факт, что распределение модификаций в открытом хроматине крайне неоднородно и зависит от функции геномного участка (энхансер, базальный промотор, 5'- либо 3'-участок транскрибируемого района). Например, у активно транскрибируемых генов модификации H3K4me3, H3K56ac, H4K16ac максимальны в окрестностях старта транскрипции, тогда как максимальные уровни модификации H3K36me3, напротив, обнаруживаются ближе к 3'-концу гена. В то же время модификации гистонов H3K4me1, H3K4me2, H3K36me2 и H3K36me3 наиболее часто встречаются в центральных участках транскриптов [100, 104].

Маркеры хроматина и транскрипционная активность генов

С помощью методов ChIP-chip, ChIP-seq и RNA-seq возможно получение полногеномных профилей локализации хроматиновых маркеров и выявление закономерностей их совместной встречаемости в различных районах геномов (промоторах, энхансерах, транскрибируемых районах, межгенных спейсерах).

В работе [104] методом ChIP-seq проанализированы 39 вариантов модификаций гистонов в районе −/+1000 пн относительно старта транскрипции 12726 генов, экспрессирующихся в CD4+ Т-клетках человека. Установлено, что ацетилирование гистонов положительно коррелирует с экспрессией генов, что хорошо согласуется с ролью этих модификаций в активации транскрипции. В то же время для генов с низкой экспрессией характерно наличие маркера H3K27me3 и отсутствие любых вариантов ацетилирования.

В исследовании, выполненном на *D. melanogaster*, проанализированы полногеномные карты распределения маркеров хроматина H3K4me3, H3K9me3, H3K27me3, H3K9ac, белка HP1a, а также РНК-полимеразы II (Pol II), полученные методом ChIP-seq у взрослых особей. Выявлены особенности распределения модификаций гистонов H3K4me3 и H3K9ac в пределах активно транскрибируемых генов: 1) наибольший уровень модификации H3K4me3 наблюдался в районе старта транскрипции; 2) модификация H3K9ac встречалась существенно чаще в центральной части транскрипта и в районе −1000 пн перед точкой терминации транскрипта [105]. Корреляции между картинами модификаций гистонов и осо-

бенностями экспрессии генов выявлены также у *A. thaliana* [106].

Предсказание транскрипционной активности генов на основе анализа характеристик хроматина

В исследованиях последних лет получены доказательства того, что модификации хроматина обеспечивают тонкую настройку количественного уровня транскрипции генов. Разработаны компьютерные методы предсказания уровня экспрессии генов на основе анализа распределения хроматиновых маркеров, а также их совместной встречаемости в различных участках генов, экспрессируемых в эмбриональных стволовых клетках мыши [107, 108], различных линиях клеток человека [109, 110], а также в клетках *C. elegans* [111].

Работа [107] на эмбриональных стволовых клетках мыши (ESC) была направлена на выяснения вопроса о соотношении двух кодов регуляции транскрипции, один из которых основан на модификациях хроматина, а второй – на расположении регуляторных мотивов и взаимодействующих с ними ТФ. В этой работе проанализированы: а) полногеномные профили экспрессии генов на основе RNA-seq и microarray; б) профили распределения модификаций семи гистонов и 12 ТФ (с помощью ChIP-seq). Рассматривались области геномной ДНК вокруг старта транскрипции и сайта терминации транскрипции каждого гена протяженностью по 8000 пн (по 4000 пн в плюс и минус направлении), подразделенные на 80 локальных участков (бинов) длиной 100 пн (рис. 4,*а*). Для каждого участка были вычислены коэффициенты корреляции (рис. 4,*б*, *в*) между уровнями экспрессии генов и характеристиками связывания с ТФ (либо частотами модификаций гистонов). Исходя из этих результатов, были построены регрессионные математические модели, позволяющие предсказывать количественный уровень экспрессии генов на основе характеристик связывания ТФ либо особенностей распределения хроматиновых маркеров. Обе модели с высокой точностью предсказывали уровень транскрипции генов (рис. 4,*г*, *д*).

Было показано также, что модели, построенные на основе характеристик белок-кодирующих генов в мышиных клетках ESC, пригодны и для предсказания уровня экспрессии генов, кодирующих microRNA в этих клетках.

Сходный подход, примененный к анализу только гистоновых маркеров в семи линиях клеток человека, а также данных экспериментов CAGE, RNA-PET и RNA-Seq о локализации стартов транскрипции генов от группы GENCODE, позволил заключить, что модели предсказания уровней экспрессии генов по характеристикам хроматина, полученным в одном клеточном типе, хорошо работают и на других типах клеток [109].

Свидетельства высокой предсказательной силы моделей, основанных на анализе распределения характеристик хроматина в эмбриональных стволовых клетках мыши, были получены и в другом исследовании [108]. Точность метода предсказания транскрипционной активности генов, основанного на данных о распределении семи типов модификаций гистонов в сочетании с данными о сайтах гиперчувствительности к ДНКазе I, была сравнимой с точностью предсказания метода, построенного на основе анализа распределения ChIP-seq пиков ТФ.

В работе Kwasnieski и соавт. [110] была исследована регуляторная активность геномных районов, для которых консорциумом ENCODE был предсказан регуляторный потенциал (энхансеры; слабые энхансеры; энхансеры, функционирующие в эмбриональных стволовых клетках). Экспериментальная проверка осуществлялась с использованием современного высокопроизводительного метода исследования репортерных конструкций CRE-seq (cis-regulatory element analysis by sequencing) и клеток K562 либо H1-hESC. Было выявлено, что существенная доля предсказанных энхансеров (26%) обладает регуляторной активностью.

Таким образом, рассмотренные выше экспериментально-теоретические исследования демонстрируют высокую предсказательную силу моделей, построенных на основе учета локализации и совместной встречаемости характеристик хроматина. Ценность этих работ заключается не только в подробном описании моделей и методов их создания, но и в том, что они с неопровержимой ясностью подтверждают значимость гистонового кода в регуляции транскрипции.

Закономерности организации хроматинового кода регуляции транскрипции

Широкомасштабное исследование закономерностей организации хроматинового кода регуляции транскрипции было проведено [112] с использованием метода максимизации энтропии на основе полногеномного анализа профилей 73 различных характеристик хроматина на клеточной линии S2 (*D. melanogaster*), полученных в рамках проекта modENCODE. В число исследуемых характеристик хроматина входили модификации гистонов, связывание с негистоновыми белками, а также варианты/субъединицы гистоновых белков, полученные на основе технологии ChIP-chip. По обучающей выборке, содержащей 265560 участков хроматина (размером 200 нуклеотидов каждый), были вычислены наблюдаемые частоты индивидуальных характеристик хро-

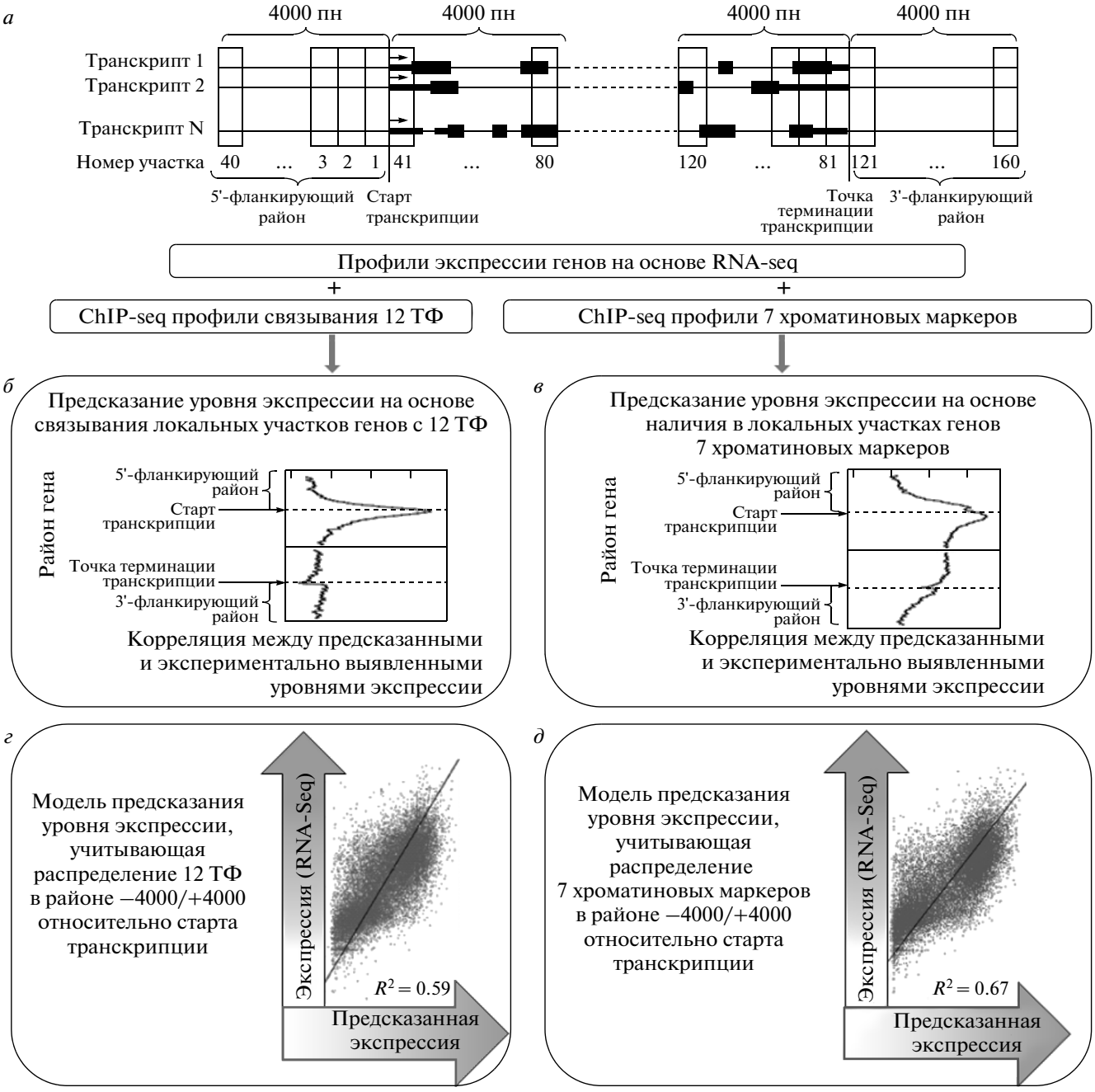


Рис. 4. Компьютерные модели предсказания транскрипционной активности генов в эмбриональных стволовых клетках мыши, построенные на основе учета особенностей характеристик связывания 12 ТФ либо распределения семи гистоновых маркеров (представлено по [107] с модификациями). *а* – схема разделения регуляторных районов генов в окрестностях стартов транскрипции и сайтов терминации транскрипции на локальные участки длиной 100 нуклеотидов и данные для построения моделей; *б*, *в* – величины коэффициентов корреляции между уровнем экспрессии генов и предсказательной силой сигналов, соответствующих 12 ТФ либо семи хроматиновым маркерам в локальных участках генов; *г*, *д* – сопоставление величин транскрипционной активности генов, рассчитанных на основе регрессионных моделей, с величинами, измеренными экспериментально с помощью RNA-Seq.

матина, а также их пар и триплетов и получены оценки их статистической значимости. Это позволило построить статистическую модель для предсказания совместной встречаемости факторов хроматина, которая включала как качественные характеристики (положительные либо отрицательные корреляции), так и количественные (вероятности совместной встречаемости). Проверка модели на независимых данных, включающих 183859 контрольных участков хроматина из

этой же линии клеток S2, продемонстрировала ее высокую предсказательную способность. Модель хорошо работала и при ее проверке на данных из другой линии клеток дрозофилы BG3, достаточно точно предсказывая индивидуальные характеристики хроматина. Это было продемонстрировано на основе анализа характеристик распределения 47 хроматиновых маркеров, изученных для линий клеток S2 и BG3.

Сходное биоинформатико-экспериментальное исследование было предпринято для обработки данных о состояниях хроматина (53 характеристики) у модельного эукариотического организма инфузории *Tetrahymena*. Для 15 различных состояний этого одноклеточного организма были идентифицированы пять вариантов функционального состояния хроматина с взаимозависимыми характеристиками, специфичными для различных клеточных процессов (стадии репликации, варианты транскрипционной активности, репарации ДНК) [113].

Результаты работ [112, 113] свидетельствуют о том, что высокоорганизованная структура хроматина накладывает существенные ограничения на разнообразие парных и тройных комбинаций гистоновых маркеров. В то же время разнообразие возможных парных и тройных комбинаций гистоновых маркеров даже в рамках этих ограничений обеспечивает огромную информационную емкость хроматинового кода для записи эпигенетической информации, значимой для экспрессии генов.

Механизмы формирования кода разметки хроматина

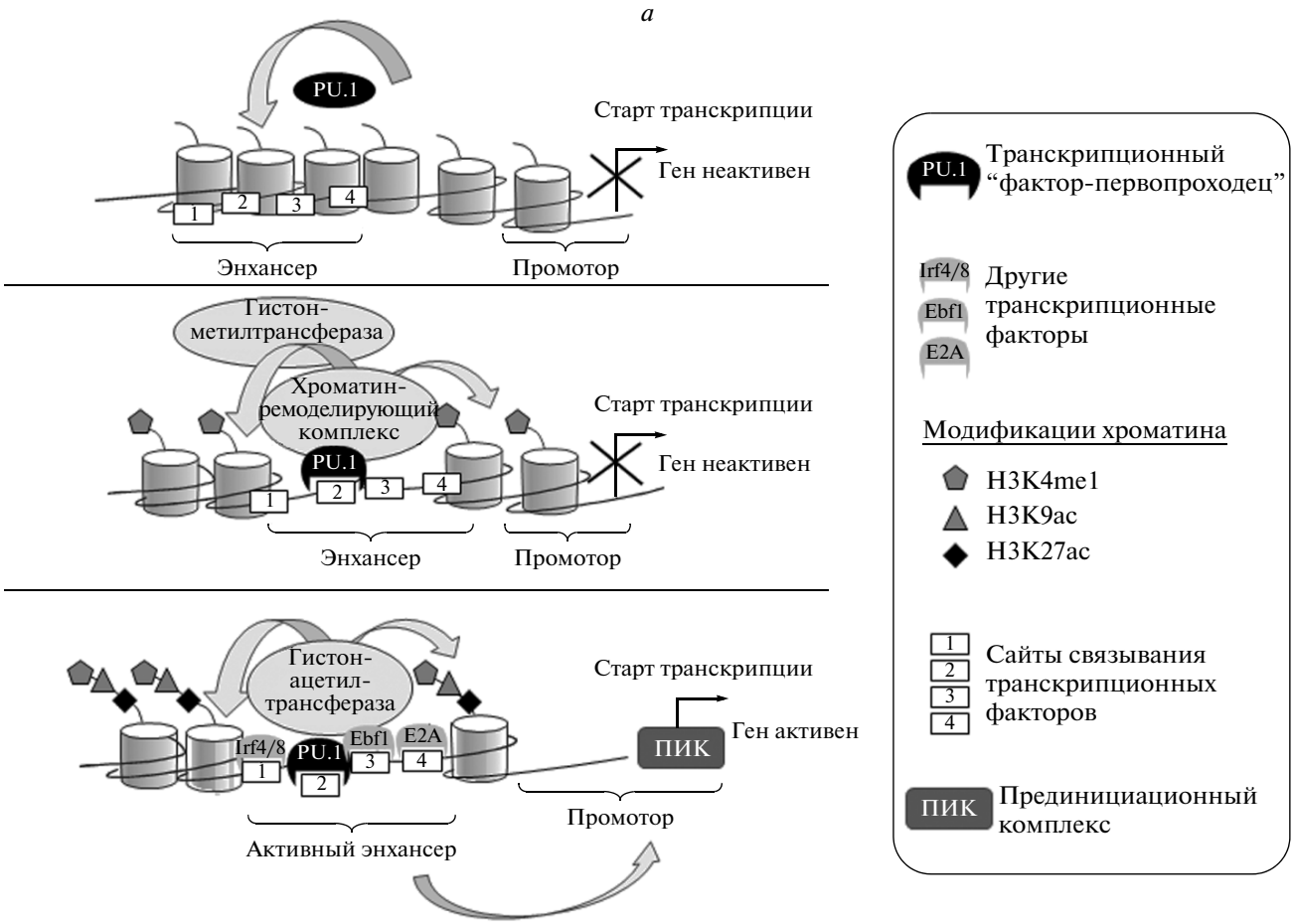
Выше указывалось, что как 5'-фланкирующие участки активно транскрибируемых генов, так и различные области транскрибируемых районов генов характеризуются определенным набором маркеров хроматина.

Состояние хроматина определяется регуляторными белками (транскрипционными корегуляторными белками). К их числу относятся белки, осуществляющие ковалентные модификации гистоновых белков, метилирование ДНК, а также АТФ-зависимое ремоделирование хроматина [114]. Согласно современной классификации, белки, регулирующие состояние хроматина, принято подразделять на три категории. К первой категории относятся так называемые белки-ридеры, считывающие информацию, записанную в форме гистонового кода. Белки-ридеры имеют в своем составе домены, способные опознавать определенные модификации хроматина. Ко второй и третьей категориям относятся белки, осуществляющие ковалентные модификации нуклеосомных белков (ацетилирование, метилирование и т.д.) и метилирование ДНК, либо возвращающие белки и ДНК в исходное состояние (деацетилазы, деметилазы). Белки второй и третьей категорий принято также называть записывающими (белки-райтеры) и стирающими (эрайзеры) белками [114].

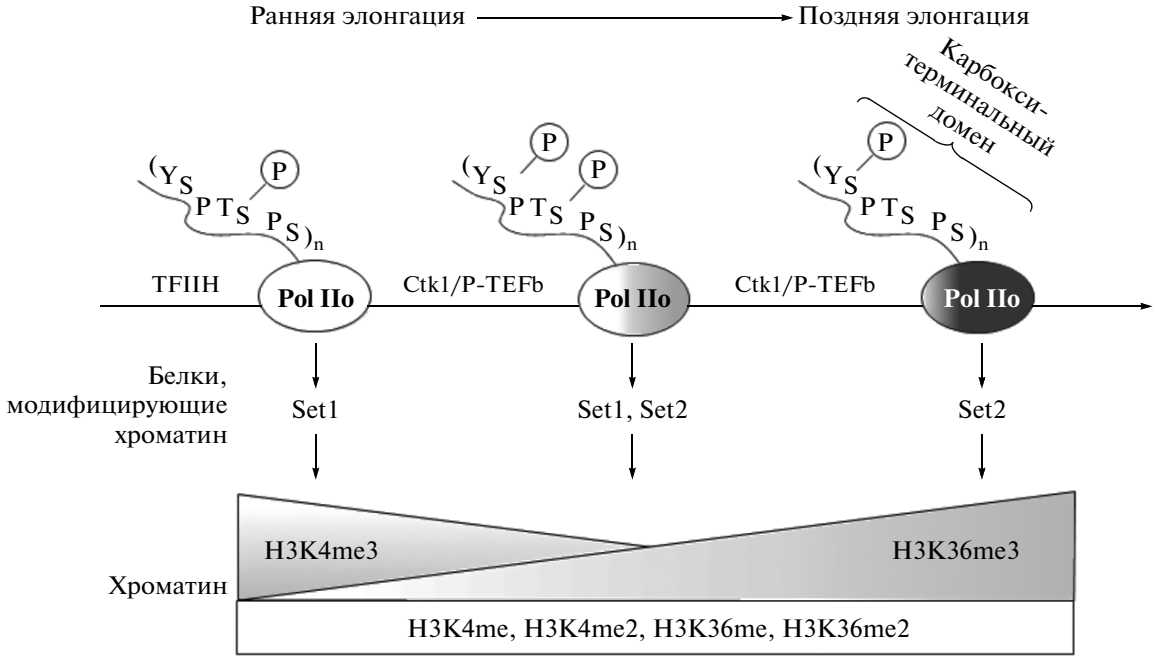
Ремоделирование хроматина и освобождение регуляторных районов гена от плотной нуклеосомной упаковки являются необходимым условием эффективной транскрипции генов эукариот, поскольку большинство ТФ не способны взаимодействовать с сайтами связывания на ДНК в составе нуклеосомы. Важную роль в процессе подготовки ДНК к взаимодействию с транскрипционной машиной (ТФ в районах энхансеров и компонентами ПИК в районе базального промотора) играют транскрипционные “факторы-первопроходцы” (pioneer transcription factors). К числу наиболее известных факторов из этой группы относятся PU.1, Pax7, FoxA1, p53 [115–117]. Последовательность событий, инициированных таким транскрипционным фактором, описана в работе [115] на примере фактора PU.1 (рис. 5,*а*). На первом этапе транскрипционный “фактор-первопроходец” PU.1 связывается со своим сайтом на ДНК в составе нуклеосомы. Далее он привлекает гистонметилтрансферазный и хроматин-ремоделирующий комплексы, которые подготавливают регуляторный район гена для последующей активации, в том числе – за счет метилирования гистонов. Затем происходит связывание с ДНК других ТФ, а также привлечение гистонацетилтрансферазы, ацетилирующей гистоны, что ослабляет плотность нуклеосомной укладки и обеспечивает возможность для взаимодействия TBP и других белков транскрипционной машины с промоторным участком гена. При этом транскрипционные “факторы-первопроходцы” могут успешно инициировать ремоделирование только в областях с относительно ослабленной нуклеосомной упаковкой, что достигается так называемой дотранскрипционной разметкой хроматина. Механизм разметки пока до конца не изучен, известно только, что частично такая разметка осуществляется за счет появления в составе нуклеосомы варианта гистонового белка H2A.Z [118].

Для транскрибируемых участков гена также характерна определенная картина нуклеосомных характеристик, обеспечивающих оптимальное взаимодействие с РНК-полимеразой. И что удивительно, сама РНК-полимераза участвует в разметке хроматина. В состав самой большой субъединицы РНК-полимеразы II входит карбокситерминальный домен (CTD), характер фосфорилирования которого меняется по мере того, как РНК-полимераза II движется от 5'- к 3'-концу гена (рис. 5,*б*). Изменяющийся характер фосфорилирования CTD определяет состав белковых комплексов, взаимодействующих с CTD на различных этапах транскрипционного цикла. На ранних

а


PU.1
Старт транскрипции
Ген неактивен
1
2
3
4
Энхансер
Промотор
Гистон-метилтрансфераза
Хроматин-ремоделирующий комплекс
PU.1
Старт транскрипции
Ген неактивен
Энхансер
Промотор
Гистон-ацетил-трансфераза
Irf4/8
PU.1
Ebf1
E2A
ПИК
Старт транскрипции
Ген активен
Промотор
Активный энхансер
Транскрипционный "фактор-первопроходец"
Irf4/8
Ebf1
E2A
Другие транскрипционные факторы
Модификации хроматина
H3K4me1
H3K9ac
H3K27ac
Сайты связывания транскрипционных факторов
ПИК
Прединициационный комплекс

б


Ранняя элонгация
Поздняя элонгация
Карбокси-терминальный домен
TFIIH
Pol IIo
Ctk1/P-TEFb
Pol IIo
Ctk1/P-TEFb
Pol IIo
Белки, модифицирующие хроматин
Set1
Set1, Set2
Set2
Хроматин
H3K4me3
H3K36me3
H3K4me, H3K4me2, H3K36me, H3K36me2

этапах транскрипции (стадия ранней элонгации) в составе белкового комплекса, взаимодействующего с РНК-полимеразой, преобладают белки семейства Set1, осуществляющие метилирование лизина в позиции H3K4. На стадии поздней элонгации, при приближении к 3'-концу гена РНК-полимераза взаимодействует преимущественно с белками семейства Set2, осуществляющими метилирование в позиции H3K36 [102, 119].

Таким образом, состояние хроматина в 5'-фланкирующих участках генов и в транскрибируемых областях генов контролируется различными механизмами, в которых задействованы различные наборы регуляторных белков.

ВЛИЯНИЕ ПОЛИМОРФИЗМА РЕГУЛЯТОРНЫХ РАЙОНОВ НА ЭКСПРЕССИЮ ГЕНОВ

Вариации связывания транскрипционных факторов с хроматином

Массовое картирование областей связывания отдельных ТФ оставляет нерешенным вопрос о том, в какой степени связывание ТФ с хроматином в определенном геномном локусе влияет на экспрессию соответствующих генов. Cusanovich и соавт. [120] исследовали влияние нокдауна 59 ТФ на экспрессию их генов-мишеней в лимфобластоидных клетках человека. Оказалось, что только небольшое количество генов, регуляторные области которых (1–10 тыс пн от старта транскрипции) связывают определенные ТФ, изменяют свою экспрессию после нокдауна этих ТФ. Это может означать, что: а) большая часть взаимодействий между ТФ и хроматином не приводит к количественно значимым изменениям уровней экспрессии генов-мишеней; б) согласно этим данным, определенная часть областей связывания ТФ, выявленных методами ChIp-seq, не функциональна. Авторы с определенными оговорками отмечают, что функционально значимое связывание ТФ сконцентрировано: а) в регуляторных элементах, которые содержат большое количество сайтов связывания ТФ; б) в сайтах с относительно высокой аффинностью связывания; в) на участках генома, которые аннотированы как “активные энхансеры” [120].

Kasowski с соавт. [121], используя комбинацию иммунопреципитации хроматина с последующим секвенированием, картировали сайты связывания Pol II и ТФ NF-κB в клетках лимфобластоидных линий, полученных от 10 индивидуумов. Межиндивидуальные различия были обнаружены в 7.5% геномных областей связывания ТФ NF-κB и 25% областей связывания Pol II, выявленных с применением очень жестких критериев. Различия связывания часто были ассоциированы с однонуклеотидными заменами или структурными геномными вариациями (CNV). Максимальное влияние на связывание ТФ NF-κB оказывали нуклеотидные замены, локализованные в районах, соответствующих консенсусной последовательности ТФ NF-κB. Оказалось, что генетические вариации, затрагивающие позиции вне консенсусных последовательностей, но в пределах пиков ChIp-seq, также влияют на связывание ТФ NF-κB или Pol II. По мнению авторов, это может объясняться кооперативными эффектами, обусловленными взаимодействиями с другими транскрипционными факторами или белками хроматина [121]. Аллельные варианты в сайтах связывания транскрипционных факторов в целом коррелируют с различиями в локальной модификации гистонов. Более того, вариации, которые влияют на хроматин в дистальных регуляторных сайтах, также часто влияют на хроматин ассоциированных с ними промоторов и экспрессию соответствующих генов [122]. С учетом выявленных межиндивидуальных различий связывания ТФ, несомненным является вклад генетической вариабельности некодирующих и межгенных районов генома в формирование фенотипических особенностей, связанных с экспрессией генов.

Нуклеотидные замены в локусах количественных признаков, ассоциированных с чувствительностью к ДНКазе I

В последнее десятилетие были разработаны методы полногеномного картирования сайтов гиперчувствительности к ДНКазе I (DHS), соответствующих областям открытого хроматина. Локализация DHS в значительной степени совпадает с ранее описанными регуляторными районами (энхансерами, промоторами, сайленсерами), а чувствительность их к ДНКазе I положительно

Рис. 5. Механизмы формирования кода разметки хроматина. *а* – пошаговая упрощенная схема ремоделирования энхансерного хроматина, индуцированного транскрипционным “фактором-первопроходцем” PU.1 (представлено по [115] с модификациями); *б* – двигаясь вдоль гена, РНК-полимераза II (Pol IIo), фосфорилированная по CTD, осуществляет транскрипционную разметку хроматина, повышая уровень его модификации и создавая благоприятные условия для последующих циклов транскрипции. На стадии ранней элонгации CTD полимеразы Pol IIo фосфорилирован преимущественно по серину в пятой позиции (S5), это фосфорилирование осуществляется киназами CDK7 и cyclin H, входящими в состав базального транскрипционного фактора TFIIH. К моменту поздней элонгации за счет действия киназы Ctk1 (у человека P-TEFb), а также S5-фосфатаз преобладающим становится фосфорилирование CTD по серину S2. Set1 и Set2 – лизинметилтрансферазы, осуществляющие метилирование гистона H3 по лизинам в 4-й и 36-й позициях (соответственно) (представлено по [119] с модификациями).

коррелирует с уровнем экспрессии генов, расположенных близко к соответствующим гиперчувствительным сайтам [123]. С учетом функциональной значимости DHS разработан метод оценки влияния нуклеотидных замен на формирование открытого хроматина, основанный на поиске dsQTL (DNase I sensitivity quantitative trait loci), т.е. локусов, для которых характерна ассоциация локальной доступности хроматина с их аллельными вариантами. Гиперчувствительные сайты, глубина прочтения DNase-seq которых достоверно коррелирует с генотипом близлежащего полиморфизма или вставки/удаления (индела), получили название dsQTL по аналогии с eQTL (expression quantitative trait loci) [124].

Поскольку dsQTL связаны со специфичными для последовательности ДНК изменениями в доступности хроматина, а часто и с изменениями в связывании ТФ, часть из них может также влиять и на уровни экспрессии близлежащих генов. Следовательно, в таких случаях dsQTL является одновременно и eQTL. Оказалось, что 55% из установленных ранее наиболее значимых eQTL были также и dsQTLs, а 39% dsQTLs были также eQTL. Кроме того, аллели более 70% совмещенных dsQTL–eQTLs были ассоциированы с доступностью хроматина и также с изменением уровня экспрессии генов [122, 124]. Следовательно, dsQTLs являются одним из важных механизмов, через который генетические вариации могут влиять на уровень экспрессии генов.

Экспериментально-компьютерные подходы к оценке эффектов нуклеотидных замен в регуляторных районах генов

Нуклеотидные замены в регуляторных районах генов могут изменять сродство транскрипционных факторов к их специфическим сайтам связывания на ДНК. Такие замены принято называть регуляторными, поскольку нарушения связывания ТФ с ДНК в регуляторных районах генов часто влекут за собой изменения в уровнях их транскрипционной активности.

По данным базы dbSNP, регуляторные районы генов человека характеризуются существенной генетической изменчивостью. Плотность однонуклеотидных замен в 5'-фланкирующих районах генов сопоставима с их плотностью в интронах и составляет 3.7 замен на 1000 нуклеотидов. С использованием данных проекта “1000 геномов” [3] была выявлена группа генов с повышенной генетической изменчивостью их 5'-фланкирующих районов (шесть замен и более на участке –500/–1), составлявшая 5.5% от общего количества генов человека [125]. Анализ с помощью системы DAVID, которая рассматривает аннотацию генов терминами из словаря Gene Ontology, показал, что эта группа обогащена генами, кодирующими ольфакторные рецепторы и белки сигнальной трансдукции, участвующие в восприятии и передаче ольфакторных (запаховых) стимулов, а также генами иммунного ответа, отвечающими за презентацию и процессинг антигенов. Повышенный уровень генетической изменчивости в промоторных районах может быть причиной вариаций в уровнях экспрессии генов системы восприятия запахов и иммунного ответа, что может объясняться необходимостью эволюционной адаптации к высоковариабельным условиям обитания, характеризующимся большим разнообразием иммуногенных и запаховых стимулов.

Для понимания молекулярных механизмов реализации эффектов регуляторных однонуклеотидных замен необходимы дальнейшие исследования, направленные на идентификацию сайтов связывания транскрипционных факторов. Огромные преимущества для решения такой задачи дает экспериментально-теоретический подход, который был описан нами ранее на примере исследований по сайтам связывания SF-1 и SREBP [126, 127]. На первом этапе экспериментально-теоретических исследований данные из научных публикаций об экспериментально-подтвержденных сайтах связывания транскрипционных факторов SF-1 и SREBP были аккумулированы в базе TRRD [64]. Далее были построены выборки нуклеотидных последовательностей сайтов SF-1 и SREBP и разработаны компьютерные методы их распознавания SITECON [128] и SiteGA [129]. С использованием данных программ были выявлены потенциальные сайты связывания SF-1 и SREBP в генах позвоночных. Точность предсказаний оценивалась с помощью экспериментов, проведенных *in vitro* методом EMSA. На следующем этапе решалась задача повышения точности предсказания сайтов SF-1. С этой целью был разработан компьютерный метод предсказания, основанный на комбинации SiteGA и оптимизированного метода весовых матриц, позволявший распознавать 80% сайтов SF-1 из контрольной выборки (недопредсказание 20%), при этом уровень ложноположительных сайтов (перепредсказание) составлял 7×10^{-5} [126]. Продолжением данной серии работ по верификации компьютерных методов распознавания сайтов связывания ТФ явилось исследование [23], в котором с использованием данных ChIP-seq по связыванию фактора FoxA2 и экспериментального метода EMSA была оценена точность распознавания сайтов связывания FoxA2 четырьмя различными компьютерными методами. Было показано, что наиболее эффективным является метод, основанный на комбинации моделей, построенных методами SiteGA и diChIPMunk/ChIPMunk.

Теоретический подход к выявлению регуляторных нуклеотидных замен, основанный на учете данных проекта ENCODE, полученных мето-

дом ChIP-seq, был предложен в работе [130]. Подход основан на предположении о том, что обогащенность геномного района пиками связывания с ТФ, выявленными с помощью ChIP-seq, свидетельствует о том, что данный район может выполнять регуляторную функцию. Это означает, что нуклеотидные замены, выявленные в пределах данного района, с наибольшей вероятностью влияют на регуляцию транскрипции. Для оценки работоспособности данного подхода была проведена экспериментальная проверка функциональной значимости однонуклеотидных замен методом задержки в геле (EMSA), подтвердившая высокую эффективность предложенного подхода.

ЗАКЛЮЧЕНИЕ

В обзоре рассмотрены ключевые направления регуляторной геномики, использующие в качестве инструмента экспериментально-компьютерные подходы. Огромные объемы данных, полученные методами NGS, с одной стороны, открывают возможность для исследования закономерностей организации регуляторных районов генов и молекулярных механизмов регуляции транскрипции, а с другой стороны требуют тесной интеграции экспериментальных исследований с теоретическими и разработки новых методов и подходов к анализу данных.

Работа выполнена при поддержке РНФ (проект № 14-24-00123).

СПИСОК ЛИТЕРАТУРЫ

1. *Pareek C.S., Smoczynski R., Tretyn A.* Sequencing technologies and genome sequencing // J. Appl. Genet. 2011. V. 52. P. 413–435.
2. *Xuan J., Yu. Y., Qing T. et al.* Next-generation sequencing in the clinic: promises and challenges // Cancer Lett. 2013. V. 340. P. 284–295.
3. *1000 Genomes Project Consortium, Abecasis G.R., Auton A.* An integrated map of genetic variation from 1,092 human genomes // Nature. 2012. V. 491. P. 56–65.
4. *Bernstein B.E., Stamatoyannopoulos J.A., Costello J.F. et al.* The NIH Roadmap epigenomics mapping consortium // Nat. Biotechnol. 2010. V. 28. P. 1045–1048.
5. *Chen G.G., Diallo A.B., Poujol R. et al.* BisQC: an operational pipeline for multiplexed bisulfite sequencing // BMC Genomics. 2014. V. 16. P. 290.
6. *Rodriguez J., Frigola J., Vendrell E. et al.* Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers // Cancer Res. 2006. V. 66. P. 8462–8468.
7. *Jacinto F.V., Ballestar T., Esteller M.* Methyl-DNA immunoprecipitation (MeDIP): Hunting down the DNA methylome // BioTechniques. 2008. V. 44. P. 35–43.
8. *Bonder M.J., Kasela S., Kals M. et al.* Genetic and epigenetic regulation of gene expression in fetal and adult human livers // BMC Genomics. 2014. V. 15. P. 860.
9. *Johnson D.S., Mortazavi A., Myers R.M., Wold B.* Genome-wide mapping of in vivo protein-DNA interactions // Science. 2007. V. 316. P. 1497–1502.
10. *Shanker A.* Genome research in the cloud // OMICS. 2012. V. 16. P. 422–428.
11. *Liu L., Li Y., Li S. et al.* Comparison of next-generation sequencing systems // J. Biomed. Biotechnol. 2012. ID. 251364.
12. *Lee H., Schatz M.C.* Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score // Bioinformatics. 2012. V. 28. P. 2097–2105.
13. *Bailey T., Krajewski P., Ladunga I. et al.* Practical guidelines for the comprehensive analysis of ChIP-seq data // PLoS Comput. Biol. 2013. V. 9. P. e1003326.
14. *Chung D., Park D., Myers K. et al.* dPeak: high resolution identification of transcription factor binding sites from PET and SET ChIP-Seq data // PLoS Comput. Biol. 2013. V. 9. P. e1003246.
15. *Lesluyes T., Johnson J., Machanick P., Bailey T.L.* Differential motif enrichment analysis of paired ChIP-seq experiments // BMC Genomics. 2014. V. 15. P. 752.
16. *Drucker T.M., Johnson S.H., Murphy S.J. et al.* BIMA V3: an aligner customized for mate pair library sequencing // Bioinformatics. 2014. V. 30. P. 1627–1629.
17. *Zhang Y., Liu T., Meyer C.A. et al.* Model-based analysis of ChIP-Seq (MACS) // Genome Biol. 2008. V. 9. P. R137.
18. *Li R., Li Y., Kristiansen K., Wang J.* SOAP: short oligonucleotide alignment program // Bioinformatics. 2008. V. 24. P. 713–714.
19. *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. P. R25.
20. *Hah N., Danko C.G., Core L. et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells // Cell. 2011. V. 145. P. 622–634.
21. *Allen M.A., Andrysik Z., Dengler V.L. et al.* Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms // Elife. 2014. V. 3. P. e02200.
22. *Kulakovskiy I., Levitsky V., Oshchepkov D. et al.* From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites // J. Bioinform. Comput. Biol. 2013. V. 11. P. 1340004.
23. *Levitsky V.G., Kulakovskiy I.V., Ershov N.I. et al.* Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data // BMC Genomics. 2014. V. 15. № 80. P. 1–12.
24. *Thurman R.E., Rynes E., Humbert R. et al.* The accessible chromatin landscape of the human genome // Nature. 2012. V. 489. P. 75–82.
25. *Matsumura H., Ito A., Saitoh H. et al.* SuperSAGE // Cell Microbiol. 2005. V. 7. P. 11–18.
26. *Ferella M., Davids B.J., Cipriano M.J. et al.* Gene expression changes during Giardia-host cell interactions

in serum-free medium // Mol. Biochem. Parasitol. 2014. V. 197. P. 21–23.

27. *Karapetyan A., Buiting C., Kuiper R.A., Coolen M.W.* Regulatory roles for long ncRNA and mRNA // Cancers. 2013. V. 5. P. 462–490.
28. *Danan M., Schwartz S., Edelheit S., Sorek R.* Transcriptome-wide discovery of circular RNAs in Archaea // Nucl. Acids Res. 2012. V. 40. P. 3131–3142.
29. *Cho S., Cho Y., Lee S. et al.* Current challenges in bacterial transcriptomics // Genomics Inform. 2013. V. 11. P. 76–82.
30. *Kawaji H., Lizio M., Itoh M. et al.* Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing // Genome Res. 2014. V. 24. P. 708–717.
31. *Kanamori-Katayama M., Itoh M., Kawaji H. et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer // Genome Res. 2011. V. 21. P. 1150–1159.
32. *Murata M., Nishiyori-Sueki H., Kojima-Ishiyama M. et al.* Detecting expressed genes using CAGE // Methods Mol. Biol. 2014. V. 1164. P. 67–85.
33. *Tucker T., Marra M., Friedman J.M.* Massively parallel sequencing: the next big thing in genetic medicine // Am. J. Hum. Genet. 2009. V. 85. P. 142–154.
34. *Mangan M.E., Williams J.M., Kuhn R.M., Lathe W.C.* 3rd. The UCSC genome browser: What every molecular biologist should know // Curr. Protoc. Mol. Biol. 2014. V. 107. P. 19.9.1–19.9.36.
35. *ENCODE Project Consortium, Birney E., Stamatoyannopoulos J.A. et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project // Nature. 2007. V. 447. P. 799–816.
36. *ENCODE Project Consortium, Bernstein B.E., Birney E. et al.* An integrated encyclopedia of DNA elements in the human genome // Nature. 2012. V. 489. P. 57–74.
37. *modENCODE Consortium, Roy S., Ernst J. et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE // Science. 2010. V. 330. P. 1787–1797.
38. *Slattery M., Ma L., Spokony R.F. et al.* Diverse patterns of genomic targeting by transcriptional regulators in *Drosophila melanogaster* // Genome Res. 2014. V. 24. P. 1224–1235.
39. *Li J.J., Huang H., Bickel P.J., Brenner S.E.* Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data // Genome Res. 2014. V. 24. P. 1086–1101.
40. *Zhimulev I.F., Zykova T.Y., Goncharov F.P. et al.* Genetic organization of interphase chromosome bands and interbands in *Drosophila melanogaster* // PLoS One. 2014. V. 9. P. e101631.
41. *Boley N., Wan K.H., Bickel P.J., Celniker S.E.* Navigating and mining modENCODE data // Methods. 2014. V. 68. P. 38–47.
42. *Kawai J., Shinagawa A., Shibata K. et al.* Functional annotation of a full-length mouse cDNA collection // Nature. 2001. V. 409. P. 685–690.
43. *Katayama S., Tomaru Y., Kasukawa T. et al.* Antisense transcription in the mammalian transcriptome // Science. 2005. V. 309. P. 1564–1566.
44. *Carninci P., Kasukawa T., Katayama S. et al.* The transcriptional landscape of the mammalian genome // Science. 2005. V. 309. P. 1559–1563.
45. *Grinchuk O.V., Jenjaroenpun P., Orlov Y.L. et al.* Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns // Nucl. Acids Res. 2010. V. 38. P. 534–547.
46. *FANTOM Consortium* and the RIKEN PMI and CLST (DGT) *et al.* Promoter-level mammalian expression atlas // Nature. 2014. V. 507. P. 462–470.
47. *Andersson R., Gebhard C., Miguel-Escalada I. et al.* An atlas of active enhancers across human cell types and tissues // Nature. 2014. V. 507. P. 455–461.
48. *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. P. 84–98.
49. *de Wit E., de Laat W.* A decade of 3C technologies: insights into nuclear organization // Genes Dev. 2012. V. 26. P. 11–24.
50. *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. P. 390–403.
51. *Niu L., Li G., Lin S.* Statistical models for detecting differential chromatin interactions mediated by a protein // PLoS One. 2014. V. 9. P. e97560.
52. *Dixon J.R., Ren B.* Topological domains in mammalian genomes identified by analysis of chromatin interactions // Nature. 2012. V. 485. P. 376–380.
53. *Lieberman-Aiden E., van Berkum N.L., Williams L. et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome // Science. 2009. V. 326. P. 289–293.
54. *Kalhor R., Tjong H., Jayathilaka N. et al.* Genome architectures revealed by tethered chromosome conformation capture and population-based modeling // Nat. Biotechnol. 2011. V. 30. P. 90–98.
55. *Баттулин Н.Р., Фишман В.С., Орлов Ю.Л. и др.* 3С-методы в исследованиях пространственной организации генома // Вавиловский журнал генетики и селекции. 2012. Т. 16. С. 872–876.
56. *Fullwood M.J., Liu M.H., Pan Y.F. et al.* An oestrogen-receptor-alpha-bound human chromatin interactome // Nature. 2009. V. 462. P. 58–64.
57. *Kieffer-Kwon K.R., Tang Z., Mathe E. et al.* Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation // Cell. 2013. V. 155. P. 1507–1520.
58. *Heintzman N.D., Hon G.C., Hawkins R.D. et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression // Nature. 2009. V. 459. P. 108–112.
59. *Gavrilov A.A., Chetverina H.V., Chermnykh E.S. et al.* Quantitative analysis of genomic element interactions by molecular colony technique // Nucl. Acids Res. 2014. V. 42. P. e36.

60. *Visel A., Rubin E.M., Pennacchio L.A.* Genomic views of distant-acting enhancers // Nature. 2009. V. 461. P. 199–205.

61. *Ghosh D.* Object-oriented transcription factors database (ooTFD) // Nucl. Acids Res. 2000. V. 28. P. 308–310.

62. *Schmid C.D., Perier R., Praz V., Bucher P.* EPD in its twentieth year: towards complete promoter coverage of selected model organisms // Nucl. Acids Res. 2006. V. 34 (Database issue). P. D82–D85.

63. *Heinemeyer T., Wingender E., Reuter I. et al.* Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL // Nucl. Acids Res. 1998. V. 26. P. 362–367.

64. *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.

65. *Suzuki A., Wakaguri H., Yamashita R. et al.* DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data // Nucl. Acids Res. 2014. Nov 5. pii: gku1080.

66. *Jiang C., Xuan Z., Zhao F., Zhang M.Q.* TRED: a transcriptional regulatory element database, new entries and other development // Nucl. Acids Res. 2007. V. 35 (Database issue). P. D137–D140.

67. *Gupta R., Bhattacharyya A., Agosto-Perez F.J. et al.* MPromDb update 2010: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-seq experimental data // Nucl. Acids Res. 2011. V. 39 (Database issue). P. D92–D97.

68. *Fernández-Suárez X.M., Rigden D.J., Galperin M.Y.* The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection // Nucl. Acids Res. 2014. V. 42 (Database issue). P. D1–D6.

69. *Zhang H.M., Chen H., Liu W. et al.* AnimalTFDB: a comprehensive animal transcription factor database // Nucl. Acids Res. 2012. V. 40 (Database issue). P. D144–D149.

70. *Wingender E., Schoeps T., Dönitz J.* TFClass: an expandable hierarchical classification of human transcription factors // Nucl. Acids Res. 2013. V. 41 (Database issue). P. D165–D170.

71. *Shipra A., Chetan K., Rao M.R.* CREMOFAC – a database of chromatin remodeling factors // Bioinformatics. 2006. V. 22. P. 2940–2944.

72. *Portales-Casamar E., Thongjuea S., Kwon A.T. et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles // Nucl. Acids Res. 2010. V. 38 (Database issue). P. D105–D110.

73. *Kulakovskiy I.V., Medvedeva Y.A., Schaefer U. et al.* HOCOMOCO: a comprehensive collection of human transcription factor binding sites models // Nucl. Acids Res. 2013. V. 41 (Database issue). D195–D202.

74. *Lenhard B., Sandelin A., Carninci P.* Metazoan promoters: emerging characteristics and insights into transcriptional regulation // Nat. Rev. Genet. 2012. V. 13. P. 233–245.

75. *Sandelin A., Carninci P., Lenhard B. et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies // Nat. Rev. Genet. 2007. V. 8. P. 424–436.

76. *Juven-Gershon T., Kadonaga J.T.* Regulation of gene expression via the core promoter and the basal transcriptional machinery // Dev. Biol. 2010. V. 339. P. 225–229.

77. *Меркулова Т.И., Ананько Е.А., Игнатьева Е.В., Колчанов Н.А.* Регуляторные коды транскрипции геномов эукариот // Генетика. 2013. Т. 49. № 1. С. 37–54.

78. *Raiber E.A., Kranaster R., Lam E. et al.* A non-canonical DNA structure is a binding motif for the transcription factor SP1 *in vitro* // Nucl. Acids Res. 2012. V. 40. P. 1499–1508.

79. *Deaton A.M., Bird A.* CpG islands and the regulation of transcription // Genes Dev. 2011. V. 25. P. 1010–1022.

80. *Smith E., Shilatifard A.* Enhancer biology and enhanceropathies // Nat. Struct. Mol. Biol. 2014. V. 21. P. 210–219.

81. *Sakabe N.J., Nobrega M.A.* Genome-wide maps of transcription regulatory elements // Wiley Interdiscip. Rev. Syst. Biol. Med. 2010. V. 2. P. 422–437.

82. *Frankel N., Davis G.K., Vargas D. et al.* Phenotypic robustness conferred by apparently redundant transcriptional enhancers // Nature. 2010. V. 466. P. 490–493.

83. *Kim T.K., Hemberg M., Gray J.M. et al.* Widespread transcription at neuronal activity-regulated enhancers // Nature. 2010. V. 465. P. 182–187.

84. *Négre N., Brown C.D., Shah P.K. et al.* A comprehensive map of insulator elements for the *Drosophila* genome // PLoS Genet. 2010. V. 6. P. e1000814.

85. *Ohlsson R., Lobanenkov V., Klenova E.* Does CTCF mediate between nuclear organization and gene expression? // Bioessays. 2010. V. 32. P. 37–50.

86. *Van Bortle K., Corces V.* tDNA insulators and the emerging role of TFIIIC in genome organization // Transcription. 2012. V. 3. P. 277–284.

87. *Ghirlando R., Giles K., Gowher H. et al.* Chromatin domains, insulators, and the regulation of gene expression // Biochim. Biophys. Acta. 2012. V. 1819. P. 644–651.

88. *Tanimoto K., Sugiura A., Omori A. et al.* A human beta-globin locus control region HS5 contains CTCF- and developmental stage-dependent enhancer-blocking activity in erythroid cells // Mol. Cell Biol. 2003. V. 23. P. 8946–8952.

89. *Hark A.T., Schoenherr C.J., Katz D.J. et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus // Nature. 2000. V. 405. P. 486–489.

90. *Lefevre P., Witham J., Lacroix C.E. et al.* The LPS-induced transcriptional upregulation of the chicken lysozyme locus involves CTCF eviction and noncoding RNA transcription // Mol. Cell. 2008. V. 32. P. 129–139.

91. *Phillips-Cremins J.E., Corces V.G.* Chromatin insulators: linking genome organization to cellular function // Mol. Cell. 2013. V. 50. P. 461–474.

92. *Hahn S., Buratowski S., Sharp F., Guarente L.* Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA se-

quences // Proc. Natl Acad. Sci. USA. 1989. V. 86. P. 5718–5722.

93. *Venters B.J., Pugh B.F.* Genomic organization of human transcription initiation complexes // Nature. 2013. V. 502. P. 53–58.
94. *Coleman R.A., Pugh B.F.* Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA // J. Biol. Chem. 1995. V. 270. P. 13850–13859.
95. *Hieb A.R., Gansen A., Böhm V., Langowski J.* The conformational state of the nucleosome entry-exit site modulates TATA box-specific TBP binding // Nucl. Acids Res. 2014. V. 42. P. 7561–7576.
96. *Пономаренко П.М., Савинкова Л.К., Драчкова И.А. и др.* Пошаговая модель связывания TBP/TATA-бокс позволяет предсказать наследственное заболевание человека по точечному полиморфизму // ДАН (биохимия, биофизика, мол. биология). 2008. Т. 419. № 6. P. 828–832.
97. *Савинкова Л.К., Пономаренко М.П., Пономаренко П.М. и др.* Полиморфизмы TATA-боксов промоторов генов человека и ассоциированные с ними наследственные патологии // Биохимия. 2009. Т. 74. С. 149–163.
98. *Savinkova L., Drachkova I., Arshinova T. et al.* An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein // PLoS One. 2013. V. 8. e54626.
99. *Drachkova I., Savinkova L., Arshinova T. et al.* The mechanism by which TATA-box polymorphisms associated with human hereditary diseases influence interactions with the TATA-binding protein // Hum. Mutat. 2014. V. 35. P. 601–608.
100. *Bannister A.J., Kouzarides T.* Regulation of chromatin by histone modifications // Cell Res. 2011. V. 21. P. 381–395.
101. *Teif V.B., Beshnova D.A., Vainshtein Y. et al.* Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development // Genome Res. 2014. V. 24. P. 1285–1295.
102. *Smolle M., Workman J.L.* Transcription-associated histone modifications and cryptic transcription // Biochim. Biophys. Acta. 2013. V. 1829. P. 84–97.
103. *Becker D., Lutsik P., Ebert P. et al.* BiQ Analyzer HiMod: an interactive software tool for high-throughput locus-specific analysis of 5-methylcytosine and its oxidized derivatives // Nucl. Acids Res. 2014. V. 42 (Web Server issue). P. W501–W507.
104. *Wang Z., Zang C., Rosenfeld J.A. et al.* Combinatorial patterns of histone acetylations and methylations in the human genome // Nat. Genet. 2008. V. 40. P. 897–903.
105. *Yin H., Sweeney S., Raha D.* A high-resolution whole-genome map of key chromatin modifications in the adult *Drosophila melanogaster* // PLoS Genet. 2011. V. 7. P. e1002380.
106. *Ha M., Ng D.W., Li W.H., Chen Z.J.* Coordinated histone modifications are associated with gene expression variation within and between species // Genome Res. 2011. V. 21. P. 590–598.
107. *Cheng C., Gerstein M.* Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells // Nucl. Acids Res. 2012. V. 40. P. 553–568.
108. *McLeay R.C., Lesluyes T., Cuellar Partida G., Bailey T.L.* Genome-wide *in silico* prediction of gene expression // Bioinformatics. 2012. V. 28. P. 2789–2796.
109. *Dong X., Greven M.C., Kundaje A. et al.* Modeling gene expression using chromatin features in various cellular contexts // Genome Biol. 2012. V. 13. P. R53.
110. *Kwasnieski J.C., Fiore C., Chaudhari H.G., Cohen B.A.* High-throughput functional testing of ENCODE segmentation predictions // Genome Res. 2014. V. 24. P. 1595–1602.
111. *Cheng C., Yan K.K., Yip K.Y. et al.* A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets // Genome Biol. 2011. V. 12. P. R15.
112. *Zhou J., Troyanskaya O.G.* Global quantitative modeling of chromatin factor interactions // PLoS Comput. Biol. 2014. V. 10. P. e1003525.
113. *Zhang C., Gao S., Molascon A.J. et al.* Bioinformatic and proteomic analysis of bulk histones reveals PTM crosstalk and chromatin features // J. Proteome Res. 2014. V. 13. P. 3330–3337.
114. *Wang Q., Huang J., Sun H. et al.* CR Cistrome: a ChIP-Seq database for chromatin regulators and histone modification linkages in human and mouse // Nucl. Acids Res. 2014. V. 42 (Database issue). P. D450–D458.
115. *Choukrallah M.A., Matthias P.* The interplay between chromatin and transcription factor networks during B cell development: Who pulls the trigger first? // Front Immunol. 2014. V. 5. P. 156.
116. *Drouin J.* Minireview: pioneer transcription factors in cell fate specification // Mol. Endocrinol. 2014. V. 28. P. 989–998.
117. *Sammons M.A., Zhu J., Drake A.M., Berger S.L.* TP53 engagement with the genome occurs in distinct local chromatin environments via pioneer factor activity // Genome Res. Publ. 2014; doi: 10.1101/gr.181883.114.
118. *Ge'vry N., Hardy S., Jacques P.E. et al.* Histone H2A.Z is essential for estrogen receptor signaling // Genes Dev. 2009. V. 23. P. 1522–1533.
119. *Cho E.J.* RNA polymerase II carboxy-terminal domain with multiple connections // Exp. Mol. Med. 2007. V. 39. P. 247–254.
120. *Cusanovich D.A., Pavlovic B., Pritchard J.K., Gilad Y.* The functional consequences of variation in transcription factor binding // PLoS Genet. 2014. V. 10. P. e1004226.
121. *Kasowski M., Grubert F., Heffelfinger C. et al.* Variation in transcription factor binding among humans // Science. 2010. V. 328. P. 232–235.
122. *McVicker G., van de Geijn B., Degner J.F. et al.* Identification of genetic variants that affect histone modifications in human cells // Science. 2013. V. 342. P. 747–749.
123. *Thurman R.E., Rynes E., Humbert R. et al.* The accessible chromatin landscape of the human genome // Nature. 2012. V. 489. P. 75–82.

124. *Degner J.F., Pai A.A., Pique-Regi R. et al.* DNase I sensitivity QTLs are a major determinant of human expression variation // Nature. 2012. V. 482. P. 390–394.
125. *Ignatieva E.V., Levitsky V.G., Yudin N.S. et al.* Genetic basis of olfactory cognition: extremely high level of DNA sequence polymorphism in promoter regions of the human olfactory receptor genes revealed using the 1000 Genomes Project dataset // Front. Psychol. 2014. V. 5. P. 247.
126. *Kolchanov N.A., Merkulova T.I., Ignatieva E.V. et al.* Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes // Brief Bioinform. 2007. V. 8. P. 266–274.
127. *Merkulova T.I., Oshchepkov D.Y., Ignatieva E.V. et al.* Bioinformatical and experimental approaches to investigation of transcription factor binding sites in vertebrate genes // Biochemistry (Mosc). 2007. V. 72. P. 1187–1193.
128. *Oshchepkov D.Y., Vityaev E.E., Grigorovich D.A. et al.* SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition // Nucl. Acids Res. 2004. V. 32. P. W208–W212.
129. *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.
130. *Bryzgalov L.O., Antontseva E.V., Matveeva M.Y. et al.* Detection of regulatory SNPs in human genome using ChIP-seq ENCODE data // PLoS One. 2013. V. 8. P. e78833.

Regulatory Genomics: Integrated Experimental and Computer Approaches

E. V. Ignatieva[a, b], O. A. Podkolodnaya[a], Yu. L. Orlov[a, b], G. V. Vasiliev[a], and N. A. Kolchanov[a, b]

[a] *Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, 630090 Russia*
[b] *Novosibirsk State University, Novosibirsk, 630090 Russia*
e-mail: eignat@bionet.nsc.ru

The review describes integrated experimental and computer approaches to the investigation of the mechanisms of transcriptional regulation of the organization of eukaryotic genes and transcription regulatory regions. These include (a) an analysis of the factors affecting the affinity of TBP (TATA-binding protein) for the TATA box; (b) research on the patterns of chromatin mark distributions and their role in the regulation of gene expression; (c) a study of 3D chromatin organization; (d) an estimation of the effects of polymorphisms on gene expression via high-resolution Chip-seq and DNase-seq techniques. It was demonstrated that integrated experimental and computer approaches are very important for the current understanding of transcription regulatory mechanisms and the structural and functional organization of the regulatory regions controlling transcription.

English translation of the paper is published in "Russian J. Genetics" (2015, vol. 51, no. 4), www.maik.ru.