ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК

На правах рукописи

Орлов Юрий Львович

ПОЛНОГЕНОМНЫЙ КОМПЬЮТЕРНЫЙ АНАЛИЗ РАСПРЕДЕЛЕНИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ ЭУКАРИОТ ПО ДАННЫМ ИММУНОПРЕЦИПИТАЦИИ ХРОМАТИНА И ВЫСОКОПРОИЗВОДИТЕЛЬНОГО СЕКВЕНИРОВАНИЯ

03.01.09 – математическая биология, биоинформатика

Диссертация на соискание
ученой степени
доктора биологических наук

Научный консультант:
академик РАН, д.б.н. Н.А. Колчанов

Новосибирск - 2014

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность проблемы

Начало XXI века ознаменовано значительными достижениями в молекулярной биологии и генетике, связанными с качественно новыми, полногеномными исследованиями. Создание высокопроизводительных автоматизированных систем секвенирования ДНК позволяет эффективно секвенировать (расшифровывать) протяженные последовательности ДНК, вплоть до целых геномов [1, 2]. Выполняются крупномасштабные проекты полного секвенирования геномов эукариот, что ведет к лавинообразному росту объема информации как о полных последовательностях геномов эукариот (http://www.ncbi.nlm.nih.gov/genbank/statistics), так и о последовательностях регуляторных районов экспрессии генов. Качественный скачок в развитии технологий массового параллельного секвенирования, таких как Roche 454, Illumina Solexa, SOLiD, за последние 5-10 лет дал импульс серии новых исследований в молекулярной биологии [2-5]. Продолжаются проекты по исследованию генетического разнообразия, полиморфизмов в популяциях [6, 7], - в настоящее время доступно более тысячи полностью секвенированных индивидуальных геномов человека. В основных молекулярно-генетических банках данных (EMBL, GenBank, DDBJ) накоплена информация о более чем 20 тысячах полностью секвенированных геномах микроорганизмов и полутора тысячах геномов эукариот, включая геном человека, причем объем расшифрованных последовательностей стремительно растет. Разработка методов высокого разрешения для анализа особенностей организации регуляторных районов генов и структуры хроматина в масштабе генома дает качественно новые данные для исследования молекулярных механизмов регуляции транскрипции генов и ставит новые задачи перед компьютерной геномикой и биоинформатикой, в том числе в проекте ENCODE [8].

В последние годы благодаря методам высокопроизводительного секвенирования ChIP-seq, ChIP-on-chip, ChIP-PET и другим ChIP-технологиям, сопряженным с иммунопреципитацией хроматина (ChIP - Chromatin ImmunoPrecipitation), появился огромный массив качественно новых данных, позволяющих оценить регуляторный потенциал клетки, в том числе исследовать все сайты связывания заданного транскрипционного фактора в геноме [9-12].

Представляемая диссертационная работа посвящена применению современных математических и компьютерных методов анализа регуляции транскрипции эукариот с

использованием данных ChIP-экспериментов, связанных с секвенированием и иммунопреципитацией хроматина.

Исследование регуляции экспрессии генов эукариот в масштабе генома требует изучения сайтов связывания транскрипционных факторов (СТТФ), контролирующих транскрипцию генов, их геномной локализации, определения генов-мишеней ТФ. Оценка числа сайтов связывания, предсказанных по нуклеотидной последовательности, только для одного транскрипционного фактора в геноме человека может достигать миллиона сайтов, что значительно превышает число генов в геноме [13]. В то же время, экспериментально установленное число сайтов варьирует от нескольких тысяч до десятков тысяч, превышая число потенциальных генов-мишеней. При этом большая часть сайтов связывания располагается в удаленных от генов районах, дистальных энхансерах, что затрудняет их компьютерное предсказание и экспериментальное исследование [3].

В последние десятилетия использовались такие подходы к определению сайтов связывания регуляторных белков, как футпринтинг ДНК, методы задержки пробы в геле (ретардация). Однако этими методами невозможно исследовать все сайты связывания транскрипционного фактора (ССТФ) в геноме. Прямое применение таких экспериментальных методов для поиска, сравнения, картирования огромного числа всех сайтов связывания, описания регуляторных районов генов в геноме невозможно из-за их большой трудоемкости и значительной стоимости.

Встают задачи исследования механизмов регуляции экспрессии генов на уровне транскрипции, связанные с развитием высокоэффективных экспериментальных методик измерения экспрессии генов, изучения динамических профилей транскрипции [4], построения карт ДНК-белковых и регуляторных взаимодействий [3]. Существующие микрочиповые технологии позволяют изучать динамику экспрессии тысяч генов одновременно [14]. Систематизация и анализ этих огромных объемов экспериментальных данных геномики и транскриптомики является сложнейшей задачей, связанной как с фундаментальными вопросами биоинформатики и системной биологии, так и с биотехнологическими приложениями, медициной, фармацевтикой.

Методы иммунопреципитации хроматина (ChIP-on-chip, ChIP-PET, ChIP-seq) с последующим массовым параллельным секвенированием позволяют исследовать сайты связывания транскрипционных факторов в масштабе генома, ставя новые задачи биоинформатики для адекватной идентификации сайтов [9, 15-18]. Исследование структуры хроматина на уровне отдельных нуклеосом (модификаций метилирования и ацетилирования гистонов в определенных позициях) с помощью технологий ChIP-seq

качественно дополняет описание регуляторных районов генов в масштабе генома [13, 19, 20]. Важным направлением исследования является построение полногеномных карт известных регуляторов плюрипотентности NANOG, OCT4, SOX2, KLF4 в стволовых клетках человека и мыши. Использование иммунопреципитации хроматина позволяет экспериментально определить контакты удаленных районов хромосом, опосредованные белковыми комплексами [21-23]. Накопилось большое количество экспериментальных данных о роли трехмерной организации генома в регуляции экспрессии генов (удаленные энхансеры, пространственные домены), полученных с помощью технологий секвенирования. Недавно появившиеся методы исследования трехмерных хромосомных контактов Hi-C [24] и ChIA-PET [12] дают качественно новую информацию о регуляторных последовательностях в геноме.

Программы анализа геномных последовательностей на персональных компьютерах стали незаменимым инструментом в экспериментальной работе молекулярных биологов. За последние десятилетия был создан широкий круг программных продуктов, направленных на изучение свойств и структуры последовательностей ДНК и белков [2, 25-28], анализа нуклеотидных последовательностей сайтов связывания, представления их в форме весовых матриц, скрытых марковских моделей, и последующего распознавания сайтов в протяженных последовательностях [27], что дает основу для теоретического компьютерного описания регуляторных районов. Большинство алгоритмов, заложенных в эти программы, применяют технику теории вероятностей и математической статистики [29], дискретной математики [30] для исследования статистических свойств и закономерностей в строении последовательностей биополимеров [27, 28]. Обработка больших объемов геномных данных требует уже использования высокопроизводительных вычислительных кластеров [28].

Важнейшей проблемой биоинформатики является проблема компьютерного исследования и поиска в геноме последовательностей, регулирующих экспрессию генов эукариот. Если раньше, в 1990-е годы, объектом исследования были одиночные последовательности и выборки последовательностей, небольшие компиляции данных и базы данных, отдельные хромосомы, и, соответственно, задачи анализа были ограничены имеющимся на тот момент объемом данных [31], то сейчас ставится задача полногеномного анализа с использованием гетерогенных интегрированных информационных ресурсов, касающихся различных аспектов организации геномов [8, 28]. К таким ресурсам, содержащим полногеномные данные, относятся базы данных экспрессии генов на микрочипах - Gene Expression Atlas [32], BioGPS [33], репозитарии

экспериментов секвенирования - GEO NCBI [34]), интегрированные средства хранения данных и визуализации геномной информации - Ensembl [35], UCSC Genome Browser [36].

Одной из ключевых задач является полногеномный компьютерный анализ распределения сайтов связывания транскрипционных факторов в геноме человека и в модельных генах эукариот по данным иммунопреципитации хроматина и высокопроизводительного секвенирования, что ставит новые задачи перед биоинформатикой, представленные в настоящей работе.

Цель и задачи исследования

Цель работы – компьютерная реконструкция структуры регуляторных районов, контролирующих транскрипцию генов эукариот на основе анализа данных о положении сайтов связывания транскрипционных факторов в геноме, полученных с помощью технологии иммунопреципитации хроматина и высокопроизводительного секвенирования (ChIP-seq).

Для достижения этой цели решались следующие задачи:

1. Разработка методов анализа данных секвенирования ChIP-seq и создание статистической модели полногеномного распределения сайтов связывания транскрипционных факторов (ССТФ).

2. Компьютерная реконструкция полногеномных карт сайтов связывания транскрипционных факторов плюрипотентности c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши. Реконструкция распределения сайтов связывания транскрипционных факторов MYC, PRDM14, ERα, FOXA1, OCT4, NANOG в геноме человека.

3. Компьютерное исследование ассоциации сайтов связывания транскрипционного фактора ERα с определенными с помощью технологии ChIP-seq маркерами хроматина, в частности, модификациями гистона H3 (H3K4me3, H3K4me1, H3K27me3, H3K9me3, H3K9ac, H3K14ac), и создание метода предсказания сайтов связывания транскрипционного фактора ERα в геноме человека на основе профилей модификаций гистонов.

4. Изучение роли хромосомных контактов в регуляции транскрипции генов человека на моделях РНК-полимеразы II и транскрипционного фактора ERα на основе компьютерного анализа полногеномных данных ChIP-seq и ChIA-PET.

Методические задачи, решавшиеся в диссертации, включали: разработку и компьютерную реализацию на языках C++ и R: (1) алгоритмов анализа полногеномных

профилей связывания транскрипционных факторов ChIP-seq; (2) алгоритмов анализа нуклеотидных последовательностей регуляторных районов, формируемых ССТФ; (3) алгоритма анализа полноты эксперимента ChIP-seq и ChIP-PET; (4) алгоритма определения кластеров ССТФ в геноме: (5) программ обработки данных экспрессии генов на микрочипах; (6) программ интеграции данных геномной аннотации расположения генов и профилей ChIP-seq; (7) программ анализа профилей ChIA-PET и ChIP-seq.

В качестве экспериментальной информации, которая была проанализирована с помощью компьютерных методов, разработанных автором диссертации, использовались данные, полученные соавторами научных публикаций Ng H.H., Kong S. Joseph R., Liu E.T., Ruan Y., Wei C.L., Lee K.L., Clarke N. с помощью методов секвенирования ДНК в Геномном институте Сингапура, а также публично доступные данные секвенирования из GEO NCBI. Автор диссертации выражает своим коллегам благодарность за предоставление этих данных.

Научная новизна

Разработаны оригинальные программы анализа распределения сайтов связывания транскрипционных факторов в геноме на основе анализа данных секвенирования сопряженного с иммунопреципитацией хроматина ChIP-seq [16, 37, 38]. С помощью этих программ построены карты связывания транскрипционных факторов c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши [3, 39-41], транскрипционных факторов c-Myc [9], ERα [13], PRDM14 в геноме человека [42], Zic3 в геноме рыбы *D.rerio* [43]. Все карты ССТФ были получены впервые.

Компьютерные программы интеграции данных о геномной локализации указанных выше ССТФ и уровнях экспрессии генов, измеренных с помощью микрочипов, позволили получить новые данные о регуляции транскрипции генов [3, 37, 44, 45]. Разработанная компьютерная база данных наборов проб микрочипов платформы Affymetrix U133, включающая оценки качества проб – однозначности картирования, соответствия целевым (таргетным) транскриптам, была новой на момент публикации, и использовалась для анализа присутствии транскриптов в цис-антисенс ориентации [46-49].

Исследование распределения нуклеосом в промоторных районах с помощью собственных компьютерных программ и анализа данных прямого секвенирования нуклеосомной ДНК дрожжей [50, 51] показало отсутствие предпочтения к

позиционированию нуклеосом *in vivo* по сравнению с данными in vitro и теоретическим предсказанием на основе контекста.

Компьютерный анализ впервые показал статистически значимую совместную локализацию сайтов связывания транскрипционных факторов Oct4, Sox2, Nanog, с одной стороны и c-Myc, n-Myc с другой, в эмбриональных стволовых клетках (ЭСК) мыши, рассчитанную по данным ChIP-seq [3, 40, 41]. Объединенные полногеномные карты расположения сайтов связывания транскрипционных факторов в геноме человека для эмбриональных стволовых клеток (ЭСК) впервые представлены в форме матриц сближенности (тепловых карт) [3, 52]. Впервые получено распределение сайтов связывания транскрипционного фактора PRDM14 в геноме для ЭСК человека и определен нуклеотидный мотив связывания [42].

Впервые построен компьютерный метод предсказания сайтов связывания ТФ ERα в масштабе генома на основе профилей модификации хроматина - ацетилирования и метилирования гистона H3 (H3K4me3, H3K4me1, H3K27me3, H3K9me3, H3K9ac, H3K14ac), определенных с помощью технологии ChIP-seq в клеточных линиях MCF-7 и T47D [13, 37]. Данные по модификациям хроматина для 16 библиотек ChIP-seq в первый раз использовались в едином компьютерном исследовании для компьютерного предсказания связывания ERα.

С помощью разработанных автором компьютерных программ карты хромосомных контактов, опосредованных связыванием белка рецептора эстрогенов ERα [21], полученные посредством технологии секвенирования парных концов ChIA-PET в клетках MCF-7, впервые проанализированы совместно с данными ChIP-seq. Впервые на основе компьютерного анализа интегрированных полногеномных данных о хромосомных контактах, опосредованных комплексами РНК-полимеразы II, сайтах связывания транскрипционных факторов, транскрипционной активности генов, и профилей модификаций гистонов для пяти клеточных линий в геноме человека показана положительная корреляция участков хромосомных контактов с модификациями гистонов, характеризующими открытое состояние хроматина (H3K4me3, H3K9ac, H3K4me1) [12].

Теоретическое значение работы. Разработанная компьютерная статистическая модель распределения сайтов связывания транскрипционных факторов позволяет достоверно определять локализацию ССТФ в геноме и оценивать полноту эксперимента по координатам секвенированных прочтений ChIP-seq.

Построена компьютерная модель, обеспечивающая высокую точность предсказания локализации сайтов связывания транскрипционного фактора - рецептора эстрогенов ERα в геноме человека за счет одновременного анализа как нуклеотидных последовательностей, так и профилей модификации хроматина (ацетилирования и метилирования гистонов), рассчитанных по данным ChIP-seq.

Представлена компьютерная модель хромосомных петель регуляторных районов транскрипции в геноме человека, опосредованных комплексом РНК-полимеразы II основанная на данных ChIA-PET.

Научно-практическая ценность разработанных методов состоит в программах анализа регуляторных районов генов по данным секвенирования в масштабе генома, полученных картах сайтов связывания сайтов связывания транскрипционных факторов Oct4, Nanog, Sox2, E2f1, n-Myc, c-Myc, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши, онкогенов MYC и ERα в геноме человека.

Программный комплекс ICGenomics (http://www-bionet.sscc.ru/icgenomics/) для исследования регуляторных районов генов и функциональной аннотации геномных последовательностей обеспечивает существенное расширение методов компьютерного анализа полногеномных данных [44]. Разработана база данных цис-антисенс транскриптов и качества проб микрочипов Affymetrix U133 [46, 47], база данных экспрессии генов на микрочипах для крыс RatDNA [53] (свидетельство госрегистрации базы данных RatDNA № 621051 от 10.10.2012г.).

Созданное Интернет-доступное программное обеспечение позволяет выполнять анализ распределения сайтов связывания транскрипционных факторов, их функциональную аннотацию (http://www-bionet.sscc.ru/icgenomics/, http://pixie.bionet.nsc.ru/ratdna/rat/index.php).

Разработан учебный курс по компьютерной геномике (Кафедра информационной биологии ФЕН НГУ), учебные материалы представлены на Школе молодых ученых по системной биологии и биоинформатике SBB-2013 (http://conf.nsc.ru/sbb2013), съезде-конференции ВОГиС-2013.

По тематике данной работы выполнены госконтракты Министерства образования и науки РФ на разработку программного обеспечения для геномных исследований (№07.514.11.4003 «Разработка алгоритмов и программных систем для решения задач анализа последовательностей, возникающих в теоретической и прикладной геномике», № 16.513.12.3107 «Проведение проблемно-ориентированных поисковых исследований в области ДНК-чипов в рамках технологической платформы «Медицина будущего»»,

№ 16.512.11.2274 «Проведение проблемно-ориентированных поисковых исследований по тематике технологической платформы "Медицина будущего" в области поиска молекулярных мишеней онкологических заболеваний с помощью биоинформационных и постгеномных технологий»), гранты РФФИ (00-04-49229-а, 01-07-90376-в, 02-07-90355-в, 03-04-48506-а, 03-04-48555-а, 03-07-90181-в, 03-07-96833-р2003югра_в, 05-04-49111-а, 05-07-90185-в, 05-07-98012-р_обь_в, 11-04-01771-а, 11-04-01888-а, 11-04-92712-ИНД_а, 12-04-00897-а, 14-04-01906), Интеграционные проекты СО РАН (119), проект 8740 Минобрнауки России «Научные и научно-педагогические кадры инновационной России» на 2009 – 2013 годы «Интегрированная биоинформационная платформа анализа данных экспрессии генов в тканях мозга», начата работа по гранту РНФ 14-14-00269.

Положения, выносимые на защиту

1) Разработанная статистическая модель полногеномного распределения сайтов связывания транскрипционного фактора позволяет оценивать полноту эксперимента по секвенированию и иммунопреципитации хроматина ChIP-seq и рассчитывать статистически значимые оценки нижней и верхней границ общего числа сайтов связывания в геноме для исследуемого фактора.

2) Полногеномные карты сайтов связывания транскрипционных факторов в эмбриональных стволовых клетках, построенные по данным ChIP-seq для c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши свидетельствуют о совместной локализации групп сайтов связывания транскрипционных факторов Oct4, Sox2, Nanog, с одной стороны, и c-Myc, n-Myc с другой.

3) Нуклеотидные последовательности, окружающие сайты связывания транскрипционного фактора Smad2 в геноме мыши, содержат специфические группы нуклеотидных мотивов, соответствующих потенциальным сайтам связывания других транскрипционных факторов. Эти мотивы различаются для сайтов связывания Smad2, найденных в эмбриональных стволовых клетках мыши при действии внешних факторов - белка Activin и ингибитора SB431542, соответственно.

4) Расположение сайтов связывания транскрипционного фактора ERα в геноме человека положительно ассоциировано с районами метилирования и ацетилирования гистонов нуклеосом H3K4me3, H3K4me1, H3K9ac и H3K14ac. Разработан компьютерный алгоритм для предсказания сайтов связывания ERα в геноме по ChIP-seq маркерам состояния хроматина; показана высокая точность предсказания с помощью этой модели.

5) Геномные области хромосомных контактов, опосредованных комплексом РНК-полимеразы II, обогащены сайтами связывания транскрипционных факторов и участками модификаций гистонов, связанными с активацией экспрессии генов.

Личный вклад автора.

Все представленные в диссертации результаты получены лично автором. Публикации, представленные в данной работе, были написаны в соавторстве. Роль автора в статьях, где он не являлся первым автором или автором для переписки, обозначена как «analyzed the data» (компьютерный и теоретический анализ данных, статистическая обработка). Специально для данного исследования автором были написаны компьютерные программы анализа ChIP-seq, статистического моделирования, сравнения геномных координат и геномной аннотации, оценки качества наборов проб микрочипов, анализа нуклеотидных контекстов, генерации базы данных цис-антисенс транскриптов, компьютерной симуляции полногеномных данных ChIP-PET, ChIP-seq и ChIA-PET.

Ключевые биоинформатические разработки по анализу наборов проб экспрессионных микрочипов Affymetrix U133 представлены в публикациях, где автор диссертации является первым автором статей (Orlov et al. 2007a; Orlov et al., 2007b; Орлов и соавт., 2011) [46, 47, 49]. База данных цис-антисенс транскриптов в геноме человека, интегрированная с расположением наборов проб Affymetrix U133, опубликована в статье (Grinchuk, ... Orlov et al., 2010) [48]. Методика анализа полноты эксперимента ChIP-seq представлена в работе (Orlov et al., 2009) [38]. Компьютерные программы, разработанные автором, и их применения описаны в работах (Orlov et al., 2012; Kuznetsov, Orlov et al., 2007; Орлов и соавт. 2012; Орлов, 2014) [16, 37, 44, 55]. Применения к анализу экспрессии мобильных элементов представлены в работе (Орлов и соавт., 2011) [49].

В статье (Joseph, Orlov et al., 2010) [13] посвященной исследованию сайтов связывания транскрипционного фактора ERα в геноме человека автор диссертации имеет равное первое авторство.

Основные результаты применения разработанных компьютерных методов для анализа распределений сайтов связывания транскрипционных факторов представлены в статьях, написанных в соавторстве. В статье (Chen, … Orlov et al., 2008) [3] автор диссертации выполнил компьютерный анализ данных и оценил полноту эксперимента ChIP-seq для 13 различных транскрипционных факторов. Программа анализа профиля ChIP-seq, разработанная автором, использовалась в исследованиях транскрипционных факторов плюрипотентности для ЭСК мыши, опубликованных в статьях [39-41] (Yuan,

… Orlov et al., 2009; Heng, … Orlov et al., 2010; Han, … Orlov et al., 2010), а также (Lee, … Orlov et al., 2011) [54].

В статье (Chia, … Orlov et al., 2010) [42] автором диссертации выполнен анализ сайтов связывания транскрипционного фактора PRDM14 и компьютерная реконструкция генных сетей плюрипотентности в ЭСК человека. В статье (Zhao, … Orlov et al., 2007) [19] с помощью разработанных компьютерных моделей исследованы полногеномные данные модификаций гистонов H3K4me3 и H3K27me3. Обобщение анализа распределений ССТФ в ЭСК человека и мыши дано в статье (Heng, Orlov, Ng, 2010) [52]. В статье (Winata, …Orlov et al., 2013) [43] представлено применение разработанных автором программ для анализа расположения ССТФ в геноме *D.rerio*, впервые выполненном на данных ChIP-seq для этого организма.

Публикация (Zeller, … Orlov et al., 2006) [9] содержит описание применения компьютерных моделей ССТФ анализа эксперимента ChIP-PET для ТФ c-Myc. В работе (Goh, Orlov et al., 2010) [51] с помощью разработанной автором компьютерной программы исследован профиль плотности нуклеосом в геноме дрожжей. В статьях (Fullwood, … Orlov et al., 2009) [21] и (Li, …Orlov et al., 2012) [12] вклад автора состоял в компьютерной обработке данных хромосомных контактов, полученных методом ChIA-PET.

Работы автора, приведенные в списке литературы и не перечисленные выше, носили методический характер, и относились к применениям разработанных алгоритмов (Орлов и соавт. 2006; Воробьева, …, Орлов и соавт. 2005; Orlov et al.. 2006; Guo, … Orlov et al., 2010; Путта, Орлов и соавт., 2011; Суслов, …. Орлов, 2013) [50, 56-59], исследованию экспрессии генов на микрочипах (Кожевникова, … Орлов, 2012; Kozhevnikova,… Orlov et al. 2013; Медведева, … Орлов, 2013) [45, 53, 60].

Апробация работы. Результаты были представлены на Пятой, Шестой, Седьмой, Восьмой и Девятой Международных Конференциях по Биоинформатике и Регуляции Структуры Генома (BGRS'06, BGRS'08, BGRS\SB-2010, BGRS\SB-2012 и BGRS\SB-2014: Новосибирск, 2006, 2008, 2010, 2012 и 2014 гг.), конференциях HUGO (2008, Хайдарабад, Индия; 2010, Монпелье, Франция; 2013, Сингапур), конференции-школе CSHL-UK – 2007 (Хинкстон, Великобритания), Конференции A-STAR 2010г. (Сингапур), Международном Симпозиуме по Биотехнологии (Москва, 2011), Школе по биоинформатике BREW-2011 (Тарту, Эстония), конференциях Постгеном-2011 (Новосибирск) и Постгеном-2012 (Казань), Конференции по интегративной Биоинформатике IB-2012 (Ханчжоу, Китай), Международном Семинаре по Системной

биологии и медицине SysPatho-2012 (Санкт-Петербург), конференциях ВОГиС-2013 (Новосибирск), МССМВ-2013 (Москва), «Нейроинформатика-2014» (Москва).

Публикации. По теме диссертации опубликовано 33 печатные работы, из них 30 – статьи в научных изданиях (журналы по списку ВАК). Включая тезисы конференций, общее число публикаций по теме диссертации - 52.

Структура и объем работы

Диссертация состоит из пяти глав: «Обзор литературы», «Модели распределения сайтов связывания транскрипционных факторов в геноме», «Карты сайтов связывания по данным ChIP-seq», «Модификации хроматина и связывание транскрипционных факторов по данным ChIP-seq», «Хромосомные контакты и регуляция транскрипции в геноме человека». Вторая глава описывает разработку методов компьютерного анализа данных ChIP-seq и анализа экспрессии генов. Третья, четвертая и пятая главы описывают применение разработанных средств для анализа ССТФ в ЭСК человека и мыши, построение полногеномных карт, анализ распределения сайтов связывания рецептора эстрогенов ERα. В четвертой главе анализ ССТФ рассмотрен в контексте структуры хроматина и модификаций гистонов, в пятой – с точки зрения хромосомных контактов. Объем диссертации составляет 343 машинописных страницы, включая 119 рисунков и 28 таблиц. Список литературы содержит 521 ссылку.

Обзор литературы (Глава 1) содержит информацию о современных исследованиях регуляторных районах транскрипции в геноме человека, включая определение сайтов связывания с помощью технологий иммунопреципитации хроматина (ChIP). Представлены исследования по регуляции экспрессии генов, связанных с образованием опухолей (ESR1, MYC, TP53), тканеспецифичной экспрессии в клеточных культурах. Описаны подходы к изучению эмбриональных стволовых клеток (ЭСК) человека и мыши, показана роль транскрипционных факторов плюрипотентности в репрограммировании. Дан обзор проблем исследования трехмерных контактов хромосом в ядре с помощью секвенирования (методы 3C, Hi-C и ChIA-PET).

Рисунок 1.1 представляет логическую взаимосвязь Глав диссертационной работы, потоков данных и видов выполненного компьютерного анализа. Из схемы, представленной на рисунке, видно, что Глава 2 «Модели распределения сайтов связывания транскрипционных факторов в геноме» представляет компьютерные модели и алгоритмы, применение которых для полногеномного анализа сайтов связывания транскрипционных факторов (ССТФ) по данным ChIP-seq описано в следующей главе «Карты сайтов связывания по данным ChIP-seq».

Рис. 1.1. Взаимосвязь глав диссертационной работы.

Как показано на рисунке, дальнейшее применение анализа данных ChIP-seq представлено в Главах «Модификации хроматина и связывание транскрипционных факторов по данным ChIP-seq» и «Хромосомные контакты и регуляция транскрипции в геноме человека», которые также связаны между собой объектом исследования - данными о ССТФ и модификациях гистонов в геноме человека.

Глава 2 «Модели распределения сайтов связывания транскрипционных факторов в геноме» содержит описание разработанных методов и компьютерных моделей распределения сайтов связывания транскрипционных факторов в эукариотическом геноме на основе анализа профилей ChIP-seq. Представлены алгоритмы анализа данных ChIP-seq о связывании транскрипционных факторов в геноме и базы микрочиповых данных по экспрессии генов, разработанные автором [3, 9, 13, 16, 38]. Представлены модели регуляторных районов транскрипции, включающие антисенс транскрипты, описан анализ качества наборов проб микрочипа Affymetrix U133 [46, 47,

49], построение базы данных цис-антисенс транскриптов [48, 61]. Показаны примеры применения компьютерного анализа экспрессии генов на микрочипах для генов крысы [45, 53]. Описаны общие средства компьютерной интеграции геномных данных, разработанные в ИЦиГ СО РАН [49, 50, 57-60], включая программный комплекс ICGenomics [44].

Глава 3 «Карты сайтов связывания по данным ChIP-seq» посвящена описанию карт сайтов связывания транскрипционных факторов построенных автором по экспериментальным данным ChIP-seq в геноме человека, в геноме мыши и в геноме *D.rerio* [9, 13, 39, 41-43, 54]. С помощью разработанных компьютерных программ обработки данных ChIP-PET и ChIP-seq были проанализированы исходные данные и определены сайты связывания транскрипционных факторов c-Myc, STAT1, FOXA1, ERα, PRDM14 [9, 13, 42] в геноме человека, а также сайты связывания транскрипционных факторов Nanog, Oct4, Sox2, Klf4, E2f1, Esrrb, CTCF, n-Myc, c-Myc, Smad1, STAT3, Tcfcp21l, Zfx, Suz12 в геноме мыши [3]. Исследовано распределение ССТФ генов, ответственных за поддержание плюрипотентности в эмбриональных стволовых клетках (ЭСК) мыши; показано существование кластеров сайтов связывания факторов Oct4-Nanog-Sox2 [3]. Представлены аналогичные кластеры связывания OCT4-NANOG-SOX2 в ЭСК в геноме человека [42, 52].

Глава 4 «Модификации хроматина и связывание транскрипционных факторов по данным ChIP-seq» содержит описание применения разработанных компьютерных методов к исследованию модификаций хроматина и связыванию транскрипционных факторов в геноме дрожжей [51, 62, 63] и в геноме человека [19]. Проанализированы полногеномные данные по модификациям гистонов (ацетилирования и метилирования гистона H3) и сайтам связывания транскрипционных факторов ERα, FOXA1 в геноме человека [13, 21, 37]. Представлен компьютерный метод предсказания сайтов связывания ERα в масштабе генома на основе профилей модификаций гистонов (H3K4me3, H3K4me1, H3K27me3, H3K9me3, H3K9ac, H3K14ac), определенных с помощью технологии ChIP-seq. Представлено обсуждение результатов в связи с продолжающимися геномными исследованиями [12].

Глава 5 «Хромосомные контакты и регуляция транскрипции в геноме человека» представляет исследование хромосомных контактов, полученных с помощью массового параллельного секвенирования нуклеотидных последовательностей контактирующих участков хромосом по методу ChIA-PET для ERα и комплекса РНК-полимеразы II в геноме человека, с помощью разработанных автором диссертации компьютерных программ [12, 21, 64]. Показана ассоциация участков хромосомных

контактов с регуляторными районами транскрипции генов и модификациями хроматина в геноме человека [12].

В Приложении даны коды программ и схемы алгоритмов, таблицы, содержащие координаты сайтов в геноме, результаты анализа кластеризации ССТФ, описание использованных компьютерных ресурсов.

Научно-практическая ценность

Практическое применение методов анализа функциональных участков (ССТФ и регуляторных районов) состоит в возможности их исследования в масштабе генома генов с использованием современных технологий массового параллельного секвенирования. Программный комплекс ICGenomics [44] качественно дополняет существующие методы анализа нуклеотидных последовательностей. Научная ценность работы связана с количественными оценками контекстной структуры геномных последовательностей в эмбриональных стволовых клетках, что позволяет уточнить молекулярные механизмы поддержания плюрипотентности и дифференцировки.

Программы и материалы, разработанные в ходе подготовки диссертации, доступны для научно-образовательных целей в Интернете на сайте ИЦиГ СО РАН по адресам: http://bioinformatics.bionet.nsc.ru/, http://www-bionet.sscc.ru/icgenomics/, http://wwwmgs.bionet.nsc.ru/mgs/programs/complexity/, http://pixie.bionet.nsc.ru/ratdna/rat/index.php, http://conf.nsc.ru/sbb2013.

Благодарности

Автор выражает глубокую признательность научному консультанту академику РАН Колчанову Н.А., сотрудникам ИЦиГ СО РАН Д.А. Афонникову и В.А. Иванисенко за помощь в подготовке работы и обсуждение научных результатов, В.А. Кузнецову за научную дискуссию на ранних этапах работы. Автор благодарен зарубежным коллегам Guoliang Li, Yijun Ruan, Ed Liu, Neil Clarke, Bing Lim, Huck-Hui Ng за позитивный опыт работы и научного общения в международном научном коллективе.

Список сокращений

БД – база данных

ИПСК – индуцированные плюрипотентные стволовые клетки

Кб – килобаза, тысяча пар нуклеотидов

Мб – мегабаза, миллион пар нуклеотидов

нт – нуклеотид

НТП – нетранслируемая последовательность

п.о. – пара оснований ДНК

ССТФ – сайты связывания транскрипционных факторов

т.п.н. – тысяча пар нуклеотидов

ТФ – транскрипционный фактор

ЭОПК – экспериментальный образец программного комплекса

ЭСК – эмбриональные стволовые клетки

Принятые англоязычные термины

3C (Chromosome Conformation Capture) – определение структуры хромосом

ChIA-PET (Chromatin Interaction Analysis by Paired-End-Tag sequencing) – метод анализа взаимодействий хроматина с помощью секвенирования парных концов

ChIP (Chromatin ImmunoPrecipitation) – иммунопреципитация хроматина

ChIP-chip – технология иммунопреципитации хроматина на микрочипе

ChIP-PET (Chromatin ImmunoPrecipitation - Paired-End-Tags) – технология иммунопреципитации хроматина с использованием парных концов ДНК

ChIP-seq – технология иммунопреципитации хроматина с последующим секвенированием

FISH – флюоресцентная гибридизация in situ

H3K14ac – модификация гистонов – ацетилированный лизин 14 гистона H3

H3K27me3 – метилированный лизин 27 гистона H3

H3K4me3 – метилированный лизин 4 гистона H3

H3K9ac – ацетилированный лизин 9 гистона H3

Hi-C – метод определения конформаций хромосом в ядре клетки

HMM (Hidden Markov models) – скрытые марковские модели

NGS (Next Generation Sequencing) – высокопроизводительное геномное секвенирование (секвенирование следующего поколения)

Глава 1. ОБЗОР ЛИТЕРАТУРЫ

Введение

В данной Главе представлен обзор литературы по современным направлениям исследований компьютерной геномики, технологиям экспериментального определения сайтов связывания транскрипционных факторов в геноме, методам анализа регуляции экспрессии генов эукариот, и соответствующим алгоритмам биоинформатики и базам данных. Глава содержит разделы, посвященные общим задачам компьютерного анализа геномных данных и проблемам компьютерного анализа данных геномного секвенирования. Описаны продолжающиеся международные проекты геномных исследований, направленные на создание аннотации функциональных элементов генома человека и основных модельных объектов - «1000 геномов», ENCODE [8] и modENCODE [65, 66], и доступные в Интернете базы данных геномной информации, включающие исходные данные секвенирования - GEO NCBI [34], GenBank [67], Ensembl [35]. Кратко представлены основные биоинформационные алгоритмы поиска гомологии, реконструкции филогенетических деревьев, статистические методы и алгоритмы предсказания сайтов связывания и регуляторных элементов в нуклеотидных последовательностях, необходимые для дальнейшего исследования.

Отдельный раздел Главы посвящен описанию молекулярных механизмов транскрипции эукариот, организации комплекса РНК-полимеразы II, регуляции транскрипции посредством белковых транскрипционных факторов. Представлена классификация регуляторных районов генов - промоторов и энхансеров, иерархическая организация регуляторных районов транскрипции генов эукариот.

В следующем разделе Главы показаны современные методы измерения экспрессии генов на уровне транскрипции: экспрессионные микрочипы (микроэрреи), секвенирование транскриптом (RNA-seq). Представлены компьютерные методы исследования регуляторных районов и сайтов связывания транскрипционных факторов, распознавания сайтов в нуклеотидных последовательностях по обучающим выборкам. Описаны базовые экспериментальные технологии и полногеномные методы определения сайтов связывания транскрипционных факторов на основе иммунопреципитации хроматина - ChIP-seq и ChIP-PET [11, 15, 17, 18].

Дано описание групп транскрипционных факторов, исследование которых важно для медицинских приложений - онкогенов при раке и факторов поддержания плюрипотентности в эмбриональных стволовых клетках. Описаны проблемы

исследования генов и поиска генов-мишеней действия кодируемых ими белков для транскрипционных факторов p53, STAT1, MYC [68], рецептора эстрогенов ERα. В связи с исследуемыми задачами освещена роль регуляции транскрипции этих генов в возникновении раковых опухолей. Дано описание задач исследования транскрипционных факторов плюрипотентности в эмбриональных стволовые клетках (для генома человека и модельных организмов), задач оптимизации репрограммирования соматических клеток, представлены имеющиеся данные по факторам OCT4, NANOG, SOX2 и ряду других.

В конце Главы представлен обзор имеющихся данных по проблеме исследования контактирующих участков хромосом в ядре клетке и регуляции транскрипции. Описаны методы определения хромосомных контактов с помощью микроскопии и флюоресцентной in situ гибридизации (FISH), а также методы определения хромосомных контактов с помощью секвенирования: 3C, Hi-C и ChIA-PET.

В заключении Главы сформулированы возникающие задачи компьютерного исследования регуляции транскрипции на основе полногеномных данных секвенирования и иммунопреципитации хроматина.

1.1. ЗАДАЧИ КОМПЬЮТЕРНОГО АНАЛИЗА ГЕНОМНЫХ ДАННЫХ

1.1.1. Международные проекты геномных исследований

Полногеномное картирование и функциональная аннотация регуляторных последовательностей генов эукариот имеют большое значение для молекулярной биологии [28]. В целом, проблема компьютерного определения функции регуляторных районов по последовательности ДНК связана с неоднозначностью кодирования генетической информации [26, 31]. Участки ДНК, с которыми связываются транскрипционные факторы, не описываются нуклеотидной последовательностью однозначно. Сайты связывания РНК-полимераз, участки начала трансляции, регуляторные районы транскрипции генов, имеют еще более сложное строение, часто состоят из нескольких блоков, находящихся на варьирующих расстояниях.

Выявление и анализ закодированных в последовательностях ДНК функциональных сигналов - сайтов связывания ТФ и регуляторных районов - требует применения более совершенных методов биоинформатики - методов распознавания образов, статистических подходов и вычислительных алгоритмов, связанных с

обработкой огромных массивов информации. Прогресс в этой области зависит от уровня накопленных молекулярно-биологических знаний, экспериментальных и компьютерных методик, и смещается в настоящее время от предсказательных компьютерных моделей в сторону полногеномного анализа экспериментальных данных, полученных методами высокопроизводительного секвенирования, микроскопии, протеомики [28, 69]. Становится все более важен статистический анализ регуляторных последовательностей в масштабе всего генома, построение интегрированных компьютерных моделей [8, 13], а не только компьютерное предсказание на основе отдельных статистических характеристик, таких как физико-химические свойства ДНК, или частоты олигонуклеотидов.

Стремительно растут темпы исследований по секвенированию геномной ДНК [5, 67] (рис. 1.2). В настоящее время доступны последовательности более 24 тысяч полностью секвенированных геномов микроорганизмов и полутора тысячах геномов эукариот (http://www.ncbi.nlm.nih.gov/genome/browse/).

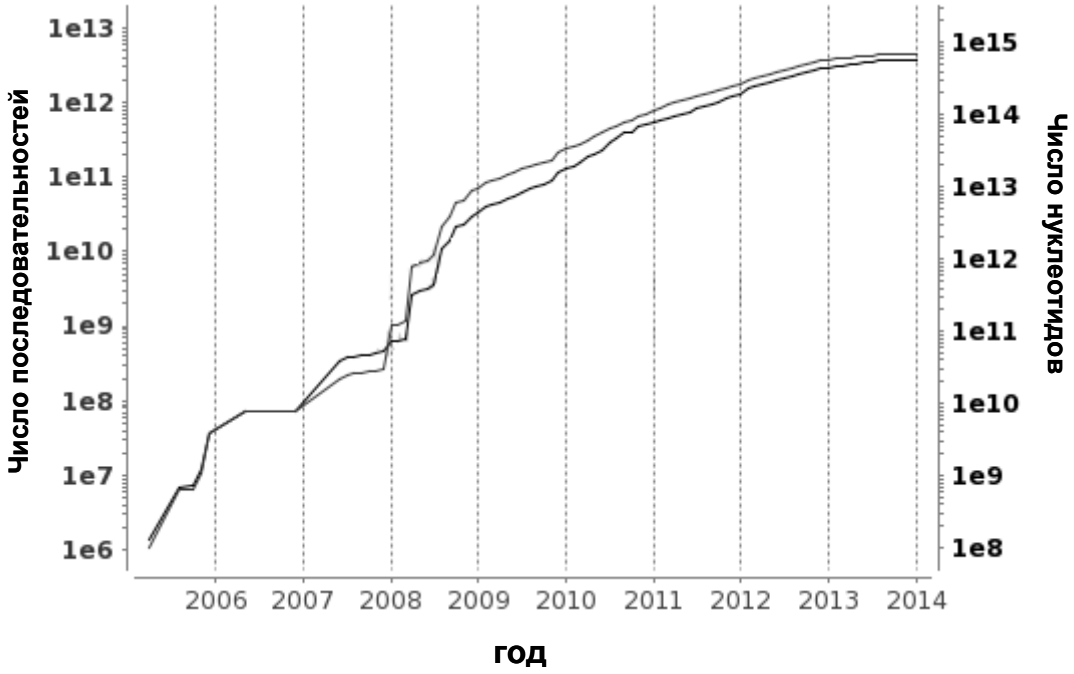


Рис. 1.2. Рост объема геномных данных в мире по информации Европейского Института Биоинформатики (начало 2014 г.), накопленной в архиве секвенированных нуклеотидных последовательностей SRA (Sequence Read Archive) (http://www.ebi.ac.uk/ena/about/statistics).
По оси Y слева - число нуклеотидных последовательностей (серый цвет) в банке данных SRA (Sequence Read Archive), справа - число нуклеотидов (черный цвет), в логарифмической шкале. Общее число нуклеотидных последовательностей в банке данных - 4.6 триллиона, число нуклеотидов - 583 триллиона.

Проект «1000 геномов» уже содержит данные о более чем тысяче индивидуальных (персональных) геномов человека. Полностью секвенированы

эукариотические геномы ряда растений, животных, включая геном мыши *Mus musculus*, геномы дрожжей *S. cerevisiae* и *S. pombe*, некоторых видов рыб, включая *Danio rerio*. Продолжается исследование геномов множества других видов, имеющих большое сельскохозяйственное или медицинское значение. Растет как общий объем геномных последовательностей, так и число видов, последовательности геномной ДНК которых представлены в международных банках данных (http://www.ncbi.nlm.nih.gov/bioproject/). Общий объем данных по нуклеотидным последовательностям удваивается каждые 2-3 года (см. рис. 1.2), а новые полностью секвенированные последовательности бактериальных геномов становятся доступными в среднем раз в две недели.

Развитие сети Интернет, появление общедоступных молекулярно-генетических баз данных большого объема предоставляет новые возможности и ставит новые задачи, связанные с исследованием полных геномов, поиском регуляторных районов транскрипции генов и их аннотации [7].

Технологии секвенирования. Отметим основные технологии высокопроизводительного секвенирования [2]: параллельное пиросеквенирование на микробусах, технология Roche 454 (www.454.com/), технология Illumina (http://www.illumina.com), использующая оптическое сканирование флюоресценции меченых нуклеотидов в клонированных колониях молекул ДНК на твердой поверхности, и технология секвенирования с помощью лигирования ABI (Applied Biosystems) (http://www.appliedbiosystems.com) SOLiD (Sequencing by Oligonucleotide Ligation). Перспективны новые технологии Ion Torrent, использующие детекцию ионов водорода во время полимеризации ДНК на гиперчувствительном сенсоре [70]. Эта компания поглощена компанией Life Technologies, которая в свою очередь поглощена компанией ThermoFisher (www.thermofisher.com). Соревнование производителей оборудования секвенирования ДНК, гонка технологий, свидетельствует о большой практической значимости полногеномных методов исследования, актуальности геномных исследований в целом.

Секвенирование ДНК на наносферах (nanoball sequencing) компании Complete Genomics (www.completegenomics.com) основано на циклической амплификации фрагментов геномной ДНК по принципу «катящегося кольца». В 2013 году компания Complete Genomics поглощена Пекинским Институтом Биоинформатики (BGI) - крупнейшим международным центром секвенирования. Компания Pacific Biosciences (PacBio) предлагает альтернативную технологию определения последовательности одиночной молекулы ДНК (технология SMRT) при считывании ДНК-полимеразой

(www.pacificbiosciences.com/). Каждая из технологий имеет свои стандарты представления данных, что требует новых компьютерных решений.

Общий тренд в технологиях секвенирования вне зависимости от физических принципов определения нуклеотидной последовательности состоит в обратной связи объемов и стоимости секвенирования, длины полученных последовательностей ДНК и требованиям к биоинформационной составляющей анализа: чем ниже цена за секвенирования за нуклеотид (за мегабазу), тем выше производительность технологии секвенирования. Чем выше производительность, тем короче получающиеся фрагменты ДНК. Чем короче прочтения ДНК (секвенированные последовательности), тем сложнее их картирование на геном, сложнее получить надежную сборку длинных последовательностей (контигов), сложнее математический аппарат и выше требования к компьютерным вычислениям. Так, секвенирование по технологии Roche 454 (www.454.com/) позволяет получать последовательности до 300 нуклеотидов, в то время как следующие технологии, такие как SOLiD, – не более 70-100 нуклеотидов и даже 35-50 нт в первых моделях. Использование более коротких последовательностей ДНК ставит технически более сложные задачи биоинформационного анализа, выравнивания, ассемблирования и картирования последовательностей.

Представление геномных последовательностей в базах данных

Гены эукариот имеют сложную «мозаичную» экзон-интронную организацию [27], включают регуляторные районы, что требует изучения с помощью современных полногеномных методов [71] и баз данных [72].

В целом, геномы эукариот характеризуются низкой плотностью кодирующих районов [73], так, для человека она составляет менее 2%. В интронах и межгенных спейсерах располагаются различные типы повторяющихся последовательностей [74], информация о которых представлена в специализированных базах данных, таких как RepeatMasker [75-77].

1.1.2. Статистические методы и алгоритмы

Классическими алгоритмами биоинформатики являются алгоритмы поиска совпадений в нуклеотидных последовательностях и выравнивания. Выравнивание последовательностей – это процедура сравнения двух (парное) или более (множественное выравнивание) последовательностей путем поиска серий (блоков) символов, находящихся в последовательностях в том же порядке [27, 78], с помощью записи их в две строки с пробелами. Метод поиска выравнивания с учетом вставок и делеций впервые был предложен в работе Нидльмана и Вунша [79]. Определяется

функция сходства F, которая учитывает число гомологичных совпадений, а также замены и вставки. В качестве параметров метода вводятся веса, увеличивающие функцию *F* при обнаружении гомологичного совпадения и штрафы за замены и вставки (делеции), уменьшающие *F*:

$$F = K_m \times V_m - K_d \times V_d - K_c \times V_c \,, \qquad (1.1)$$

где K_m, K_d, K_c – количество совпадений (*m*), делеций (*d*) и замен (*c*), V_m, V_d, V_c – параметры, характеризующие веса совпадений, делеций и замен, соответственно.

Для двух сравниваемых последовательностей строится точечная матрица гомологии. Задача решается с помощью динамического программирования. Для построения оптимального пути, соответствующего максимуму функции сходства, заполняется матрица наилучших значений функции *F(i,j)*, по которой восстанавливается весь путь в точечной матрице гомологии, соответствующий выравниванию:

$$F(i,j) = \max\{F(i-1,j) - V_d, F(i,j-1) - V_d, F(i-1,j-1) - V_{ij}\} \qquad (1.2)$$

Здесь $V_{ij} = \begin{cases} V_m & \text{если } i, j \text{ совпадают;} \\ -V_c & \text{в противном случае.} \end{cases}$

Был разработан ряд эвристических методов, имеющих в своей основе идеи метода выравнивания Нидльмана-Вунша [80], и широко применяющихся для массового анализа данных секвенирования. Для оптимизации сравнения нуклеотидных последовательностей разработано несколько алгоритмических компьютерных приемов. Метод быстрого поиска повторов в тексте, предложенный в работе [81], получил название метода *l*-граммного разложения. Такое технологическое решение применяется как для быстрого поиска гомологий в банках данных, так и для поиска повторов в протяженных последовательностях [82, 83]. Другой подход, используемый для быстрого поиска гомологий – суффиксные деревья. Этот метод может использоваться как для быстрого поиска гомологий, так и для выявления консервативных мотивов в выборках функциональных последовательностей [84].

Поиск гомологий на основе алгоритмов выравнивания FASTA и BLAST

Наиболее распространенными для массового анализа являются программы FASTA и BLAST. Программа FASTA (FAST Alignment and Search Tool - all), основанная на динамическом поиске совпадений, была разработана для быстрого поиска гомологий между двумя белковыми или нуклеотидными последовательностями [85]. Полностью совпадающие (инициирующие) участки расширяются с учетом возможных несовпадений, вставок и делеций. Далее алгоритм выполняет выравнивание методом Нидльмана-Вунша. Разработаны скоростные варианты FASTA для

картирования коротких последовательностей высокопроизводительного секвенирования (http://drfast.sourceforge.net/; http://mrsfast.sourceforge.net/)

Алгоритм BLAST (Basic Local Alignment Search Tool) работает на порядок быстрее FASTA [78]. В основе метода лежит понятие пары сегментов с высоким счетом (HSP – high-scoring segment pair), т.е. такие участков одинаковой длины, для которых получено значение функции сходства больше некоторого порогового значения. Программа BLAST (http://blast.ncbi.nlm.nih.gov/) является мировым стандартом, алгоритм имеет огромное число цитирований [69]. Программа может быть инсталлирована локально и настроена на заданные пользователем геномные базы данных. В BLAST для поиска гомологии в базах белковых последовательностей могут использоваться алгоритмы BLASTP (парное выравнивание последовательностей), PSI-BLAST (позиционно специфический итерационный BLAST), PHI-BLAST (поиск паттернов, инициируемый BLAST). В PSI-BLAST для выравнивания используются позиционно специфические матрицы весов (PSSM).

Кроме стандартных алгоритмов разработан ряд оптимизированных программ быстрого поиска протяженных последовательностей, такие как BLAT (BLAST-like alignment tool) (http://genome.ucsc.edu/cgi-bin/hgBlat) [86], картирования коротких последовательностей, реализации алгоритмов поиска для параллельной компьютерной архитектуры [2, 87].

Реконструкция деревьев сходства

Набор выровненных нуклеотидных или аминокислотных последовательностей может быть использован для восстановления (построения) филогенетического дерева. Филогенетическое дерево – это бинарный (древовидный) граф, отражающий гипотетическую картину дивергенции последовательностей. Среди методов построения деревьев можно выделить группу матричных методов [88], метод объединения соседей [89], метод максимальной экономии, метод максимального правдоподобия и некоторые другие [27, 90].

Существуют сотни пакетов программ, предназначенных для выполнения той или иной части филогенетического анализа (на 2013 год было представлено около 400 программ на http://evolution.genetics.washington.edu/phylip/software.html). Отметим, в частности, пакеты программ филогенетического анализа MEGA [91], PAML [92] и пакет VOSTORG, разработанный в ИЦиГ СО РАН [93, 94]. Распространен пакет программ PHYLIP [95], использующийся для анализа матриц расстояний между последовательностями и построения филогенетических деревьев. Разработаны

конвейеры программ для анализа эволюционных расстояний, такие как SAMEM (http://pixie.bionet.nsc.ru/samem/) [96].

Сравнение точности методов распознавания

Для сравнения методов предсказания (распознавания) функциональных элементов ДНК используются меры точности распознавания, пришедшие из математических дисциплин [97]. В терминологии теории распознавания образов функциональные районы (промоторы, сайты связывания и т.д.) соответствуют классу «Да», участки ДНК, не выполняющие такой функции, – классу «Нет». Таблица сопряженности результатов предсказания размера 2 × 2 содержит стандартную классификацию и терминологию, используемую при сравнении точности предсказания.

Таблица 1.1

		Предсказано	
Класс объектов	Число объектов	Предсказано как позитивные	Предсказано как негативные
Позитивная выборка (Класс «Да»)	TP+FN	TP	FN
Негативная выборка (Класс «Нет»)	FP+TN	FP	TN
Всего объектов	TP+FN+FP+TN	TP+FP	FN+TN

Значения из таблицы сопряженности используются для подсчетов величин, характеризующих статистическую значимость (величина p) получения наблюдаемого результата. Для оценки точности предсказания используют ошибки первого и второго рода. Ошибка первого рода E_1 – недопредсказание, доля ложно предсказанных объектов класса «Да», E_1=FP/(FP+TP). Ошибка второго рода – перепредсказание, доля ложно предсказанных объектов класса «Нет» E_2=FN/(TN+FN). Статистическая значимость может быть оценена по точному критерию Фишера [98].

Другие меры оценки точности методов предсказания – чувствительность *Sn* и специфичность *Sp* [88]. Чувствительность – доля правильных предсказаний среди всех реальных объектов, *Sn*=TP/(TP+FN), а специфичность – доля правильных предсказаний по отношению ко всем полученным предсказаниям *Sp*=TP/(TP+FP). Специфичность – обратная величина к ошибке первого рода, *Sp*=1-E_1.

Чувствительность *Sn* также называют долей истинных положительных классификаций TPR (True Positive Rate). Специфичность *Sp* классификации также называют долей ложных положительных классификаций FPR (False Positive Rate).

Для сравнения методов распознавания используют как ошибки первого и второго рода, так и чувствительность и специфичность. Для сравнения точности методов по

одному параметру в единой шкале (от -1 до +1) может использоваться корреляционный коэффициент ошибок:

$$CC = \frac{TP * TN - FN * FP}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}} \quad (1.3)$$

Площадь под кривой ошибок

Стандартом оценки точности распознавания/предсказания в биоинформатике при варьировании порога является площадь под кривой ошибок (ROC-AUC), которая принимает значения в интервале <0;1> [51, 99]. Кривая ошибок или ROC-кривая – график, позволяющий оценить качество бинарной классификации, отображает зависимость доли верных положительных классификаций (TPR) от доли ложных положительных классификаций (FPR) при варьировании порога решающего правила. Термин ROC (Receiver Operating Characteristic) - операционная характеристика приёмника пришёл из теории обработки сигналов. ROC-кривая показывает зависимость TPR от FPR при варьировании порога распознавания (Рис. 1.3).

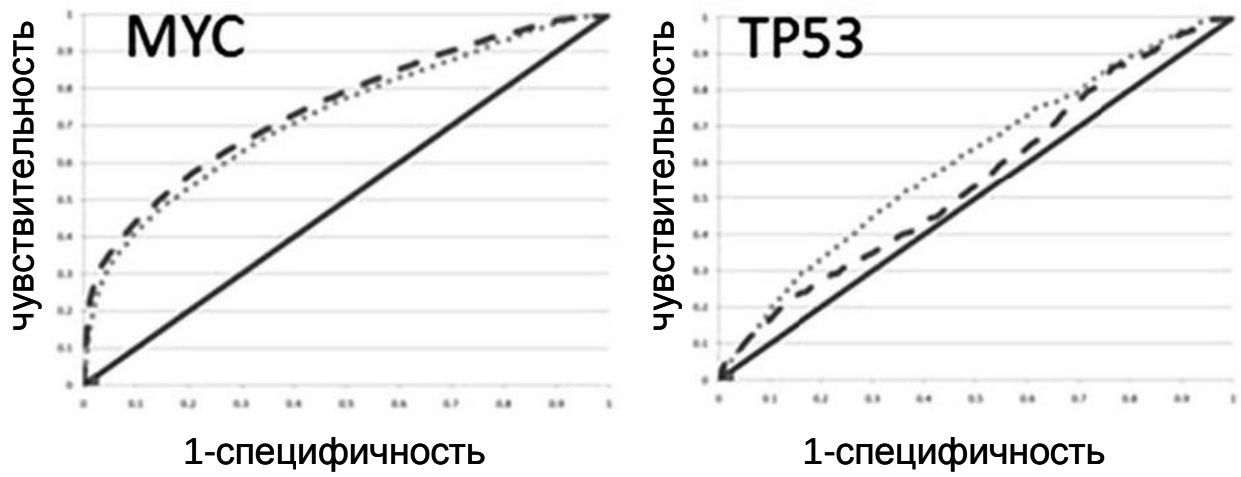


Рис. 1.3. Пример кривой ошибок для предсказания сайтов связывания транскрипционных факторов MYC и p53 (TP53) в геноме человека [99].

Кривая проходит из точки (0,0), соответствующей максимальному значению порога (все объекты классифицируются как отрицательные, и ошибки возникают на всех положительных объектах, FPR=0, TPR=0), в точку (1,1), соответствующую минимальному значению порога (все объекты классифицируются как положительные, FPR=1, TPR=1). Случайное распознавание соответствует прямой линии, площадь под кривой равна 0.5. Лучший вариант — это кривая, проходящая на графике через точки (0,0); (0,1); (1,1).

На рисунке 1.3 приведен пример кривой ошибок для предсказания сайтов связывания транскрипционных факторов в геноме человека [99]. Видно, что кривая ошибок выше для сайтов связывания MYC (левая панель), чем для сайтов связывания p53 (правая панель), и, соответственно распознавание сайтов MYC точнее [99].

Площадь под ROC-кривой AUC (Area Under Curve) является агрегированной характеристикой качества классификации, не зависящей от соотношения ошибок. Чем больше значение AUC, тем «лучше» модель. Данный показатель используется для сравнительного анализа нескольких моделей классификации. Используется также бутстреп (bootstrap) анализ, когда составляется обучающий набор из случайно выбранных элементов выборки и проверка предсказания на оставшихся данных. Такой анализ будет представлен в Главе 4 данной работы.

Для распознавания и классификации нуклеотидных последовательностей на основе набора характеристик (контекстных, физико-химических или полногеномных) используются регрессионные модели, в том числе логистическая регрессия. Обобщённые линейные модели, называемые также обобщёнными аддитивными моделями, можно рассматривать как обобщение криволинейной регрессии. Логистическая регрессия — частный случай обобщённой линейной модели, если взять логит-функцию связи $P=1/(1+e^{-y})$, где y является линейной комбинацией независимых переменных $y=w_1x_1+w_2x_2+...+w_0$. С помощью анализа зависимой переменной, принимающей значения от 0 до 1 (имеющей смысл вероятности) можно оценить ошибки классификации. Логистическая регрессия применяется для решения задач классификации и позволяет оценивать вероятности принадлежности объекта рассматриваемым классам.

Вычислительные процедуры в масштабе генома

Отметим, что оценка параметров сложных распределений, симуляция предсказаний для случайных последовательностей в геномах требует сложных вычислительных процедур. Так при генерации псевдослучайных чисел, соответствующих например позициям в нуклеотидной последовательности генома, необходим датчик чисел, избегающий повторов и периодичностей, такой как «Mersenne Twister» [100]. Виртуальное пространство для генерации распределения позиций сайтов в геноме человека (соответствующее линейным) позициям на хромосомах, составляет около 3 гигабаз - значительно больше, чем может произвести датчик случайных чисел без повторов. Датчик случайных чисел на компиляторе UNIX C++ может дать периодичность в многократно повторяемых симуляциях (поскольку рассчитан на

генерацию до 2^{31}=~2Гб, что меньше моделируемого пространства). Соответствие качества работы генерации случайных чисел без повторов этого датчика для моделирования биологических систем было показано в работе [101].

Статистический и комбинаторный анализ нуклеотидных последовательностей

Изучение нуклеотидных слов - коротких последовательностей, позволяет анализировать регуляторные последовательности генов, выдвигать гипотезы о функциональной роли отдельных фрагментов генетического текста [71, 102], в том числе регуляторных районах генов. Исторически одним из первых представление о генетических языках, содержащих нуклеотидные слова, было введено В.А. Ратнером в 1970-х [103]. Плодотворным оказался подход, связанный с лингвистическими представлениями текста и математической теорией кодирования информации [104-107]. В связи с исследованиями структуры генетических текстов Э.Н. Трифоновым разрабатывалась теория множественности кодов, содержащихся в генетических текстах [104, 108]. В первичной структуре белка представлена информация о его пространственной структуре и локализации функциональных сайтов. В первичной структуре мРНК, помимо информации о кодируемой аминокислотной последовательности, присутствует информация о вторичной структуре [109, 110]. На уровне гена, кодирующего эту мРНК, есть информация о локальной конформации ДНК в виде взаимного расположения пуриновых и пиримидиновых пар, а также информация о локализации нуклеосом – в виде участков специфического связывания с гистонами [104, 111, 112]. Таким образом, в пределах природного генетического текста может быть записано несколько генетических сообщений, определяющих различные аспекты структурно-функциональной организации макромолекул [110, 112]. Одновременная запись возможна лишь в случае, если эти генетические сообщения совместимы [113]. Так, триплетный код допускает наложение нескольких слабо позиционированных структурных сигналов в последовательности путем синонимичных замен [110, 112].

Важной универсальной характеристикой геномных последовательностей, является сложность текста [114, 115]. Интерес представляет оценка сложности генерации (порождения) текста в виде минимального числа операций копирования, необходимых для воспроизведения последовательности по ней самой (по методу Лемпеля и Зива). Такие операционные меры, адаптированные к последовательностям ДНК, были предложены В.Д. Гусевым и соавторами [115, 116], и развиты в работе автора диссертации [114]. Сложность текста может быть определена различными способами, основанными на алгоритмических оценках [105, 106], оценках энтропии

Шеннона [117], разнообразии словаря различных слов длины k (k-мер) [118]. Поиск участков низкой сложности в нуклеотидных последовательностях связан не только с теоретическими оценками распространенности повторов в геноме, но и с задачами анализа результатов высокопроизводительного геномного секвенирования, оценок уникальности картирования коротких последовательностей в геноме [57]. Разработанные автором компьютерные программы оценки сложности текста [114] применялись в данной работе для фильтрации данных ChIP-seq, анализа ошибок в прочтениях ДНК при секвенировании.

1.2 ТРАНСКРИПЦИЯ ГЕНОВ ЭУКАРИОТ

1.2.1. Транскрипция и транскрипционные факторы

Первая стадия считывания генетической информации - транскрипция ДНК проходит с участием РНК-полимераз и зависит от других белков – факторов транскрипции [73]. Синтез РНК, не зависящий от присутствия регуляторных молекул, называют базальной транскрипцией. Полногеномные методы исследования [119, 120] показали, что связанные с проксимальными промоторами РНК-полимеразы находятся одновременно примерно на 30% генов в геноме человека.

Регуляция работы генов в клетках эукариот координируется с помощью белковых факторов в зависимости от типа ткани, стадии развития организма, фазы клеточного цикла [73, 121]. Экспрессия генов эукариот связана с особенностями нуклеосомной упаковки хроматина, метилированием ДНК, интенсивностью сплайсинга, полиаденилирования, стабильностью мРНК в цитоплазме, посттрансляционными модификациями, внутриклеточным транспортом и скоростью деградации белка [122-125]. Ключевая роль в регуляции экспрессии генов принадлежит транскрипции, запускающей цепочку молекулярных процессов [126, 127]. Ядерные белковые факторы транскрипции способны связываться с последовательностями ДНК, оказывая влияние на эффективность экспрессии генов, расположенных в разных участках генома [128]. В современных исследованиях встают задачи определения сайтов связывания транскрипционных факторов в масштабе генома, определения на этой основе генов-мишеней и реконструкции регуляторных генных сетей.

ДНК-белковые контакты включают водородные связи и Ван-дер-Ваальсовы взаимодействия между боковыми группами аминокислотных остатков, расположенными на поверхности белка, и атомами азотистых оснований сайта ДНК

[128]. Разработана структурная классификация ДНК-связывающих доменов, содержащая классы основных факторов (в том числе «лейциновая застежка»), класс координированных цинком ДНК-связывающих доменов, класс «спираль-поворот-спираль» (включая гомеодомен) и класс бета-укладки с контактами по малой бороздке ДНК [129]. Большинство транскрипционных факторов контактируют со своим сайтом связывания по большой бороздке двойной спирали ДНК.

Основная (базальная) транскрипционная машина включает в себя РНК-полимеразу II и белковые комплексы - основные факторы транскрипции (GTF - от английского General Transcription Factors) TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH и TFIIK (рис. 1.5). Сборка инициирующего комплекса начинается со связывания фактора транскрипции TFIID с TATA-боксом. Сначала с TATA-боксом связывается одна из субъединиц TFIID - TATA-связывающий белок (или TBP - TATA-box binding protein) [130-132]. Белки - модуляторы транскрипции могут взаимодействовать с другими общими транскрипционными факторами [133, 134].

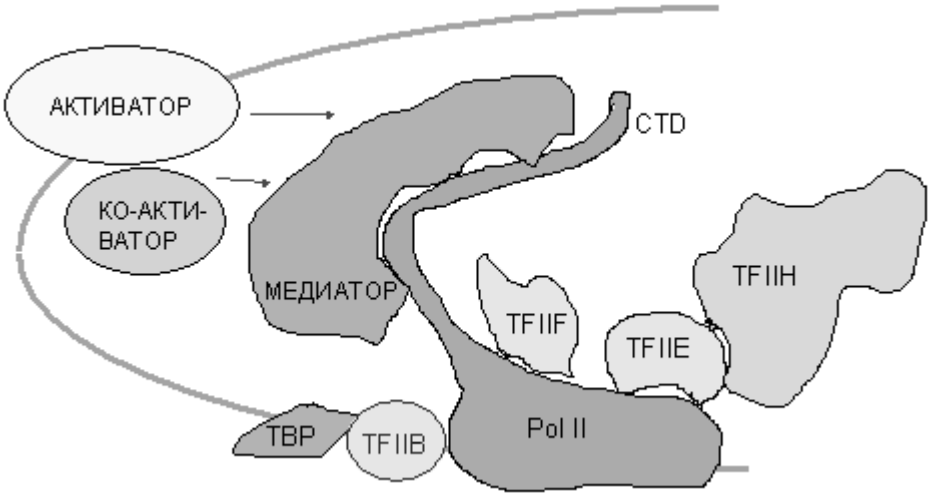


Рис. 1.4. Схема активирующих взаимодействий между активаторами, ко-активаторами и медиатором [135].

Медиатор транскрипции - это комплекс белков, связывающийся с С-концевым доменом (CTD) РНК-полимеразы II и образующий с ним полный фермент - холофермент (holoenzyme) [136]. Медиатор необходим для связи между РНК-полимеразой II и белками-активаторами транскрипции [137]. Существует ряд белков - регуляторов транскрипции, или ко-факторов, которые не контактируют непосредственно с ДНК, а взаимодействуют с другими факторами, связанными с ДНК, и могут быть либо ко-активаторами, либо ко-репрессорами.

Исследование и описание взаимодействий ко-активаторов с ДНК в регуляторных районах генов в масштабе генома представляет собой важную фундаментальную

научную проблему, для решения которой необходим компьютерный анализ полногеномных экспериментальных данных.

Комплекс РНК-полимеразы II

Существует три типа эукариотических РНК-полимераз: I, II и III. РНК-полимераза II транскрибирует белок-кодирующие гены [138]. При выходе из клеточного ядра молекулы мРНК, транскрибированные полимеразой II, проходят серию ковалентных модификаций, определяющих их функциональную специализацию и отличающих их от транскриптов, синтезированных другими РНК-полимеразами.

ТАТА-связывающий белок (TBP) – ключевой элемент механизма инициации транскрипции эукариот, входящий в состав комплекса TFIID, также необходим для осуществления транскрипции РНК-полимеразами I и III [126]. Комплекс TFIID - общий фактор транскрипции, состоящий из нескольких субъединиц, который связываясь с промотором обеспечивает формирование инициаторного комплекса [139, 140]. В состав TFIID входят до 12 факторов ТАФ (TBP-ассоциированные факторы, или TAF - TBP-associated factors).

Исследование комплекса РНК-полимеразы II, анализ распределения такого связывания в геноме важны для исследования промоторных районов белок-кодирующих генов.

1.2.2. Методы измерения экспрессии генов

Большое значение для полногеномного анализа имеет интеграция полногеномных данных по экспрессии генов (транскриптомные данные) с данными по расположению генов на хромосомах, их характеристикам, паттернам экспрессии в тканях организма [141]. Так, анализ групп высокоэкспрессирующихся генов показал, что они имеют меньшую длину и кодируются меньшим количеством экзонов [142, 143]. Количественные оценки экспрессии генов эукариот должны опираться на современные полногеномные методы.

Измерение экспрессии гена на уровне транскрипции (количество транскрибированной мРНК) может быть выполнено с помощью ПЦР в реальном времени, с помощью экспрессионных микрочипов [144], с помощью технологий EST (аббревиатура от Expressed Sequence Tags), SAGE (Serial Analysis of Gene Expression) [145, 146]. За последние годы было предложено несколько технологий анализа экспрессии генов как с помощью экзонных микрочипов, так и с помощью тотального секвенирования (RNA-seq) [4]. Разработаны карты транскриптом для различных тканей человека [141, 147].

Микрочипы получили большое распространение несколько лет назад, но в настоящее время технология микрочипов отходит, уступая по эффективности технологиям секвенирования RNA-seq. Тем не менее, накопленный за последние годы значительный массив экспериментальных, прежде всего клинических, данных об экспрессии генов на микрочипах, делает необходимым разработку оптимальных компьютерных методов для использования таких данных.

ДНК-микрочип (микропластина, или микроэррэй, от англ. - microarray) — это комплексная технология, используемая в молекулярной биологии и медицине. Микрочип состоит из нескольких (от десятков до тысяч) микроскопических ячеек на пластинке (чипе) содержащих дезокси-олигонуклеотиды. Каждая ячейка содержит ДНК специфической последовательности, которая используется для гибридизации с кДНК или мРНК. Эксперимент проводится во многих ячейках одновременно для заданного множества транскрибирующихся последовательностей (проб). Гибридизация зонда и мишени регистрируется и количественно определяется при помощи флюоресценции или хемилюминесценции.

Микрочипы отличаются по конструкции, особенностям работы, эффективности, технологическим подходам. Обычно, в микрочипе зонды ковалентно прикрепляются к твердой поверхности — стеклянному или кремниевому чипу. Распространены микрочипы компаний Affymetrix (www.affymetrix.com/), Illumina (www.illumina.com), Agilent (http://www.home.agilent.com/), NimbleGen (www.nimblegen.com/), CodeLink (www.appliedmicroarrays.com). Некоторые платформы, например, выпускаемые компанией Illumina, используют микроскопические шарики вместо твердых поверхностей. Отметим различие технологических платформ: одноцветовой микрочип (one-color) компании Affymetrix [148] и двухцветовые чипы (two-color) компаний NimbleGen и Agilent. Методы измерения уровней экспрессии генов на основе таких микрочипов получили широкое распространение в медицинских исследованиях [149-152]. Отмечена низкая корреляция между измеренной экспрессией одних и тех же генов на микрочипах и с помощью других технологий [146].

Основной статистической задачей обработки данных экспериментов на микрочипах является определение дифференциально экспрессирующихся генов. Разработан ряд пакетов для решения этой задачи, таких как SAM (Statistical Analysis of Microarrays) [153]. Не менее важны и задачи процессинга данных, адекватного определения сигнала проб на микрочипе, рассмотренные в настоящей работе.

Распространенные коммерческие платформы микрочипов (в частности Affymetrix) имеют ряд технических недостатков, связанных с несоответствием проб и

генов, для измерения транскрипции которых предназначены эти пробы [46]. Технология синтеза коротких олигонуклеотидных зондов (25 п.н.) непосредственно на поверхности микрочипа in situ с использованием литографических масок была разработана компанией «Аффиметрикс» (Affymetrix, www.affymetrix.com/) для изготовления микрочипов GeneChip. Исходно до 2003 г. был разработан микрочип GeneChip U133A, дополненный позднее чипами U133B и U133 plus 2, более полно соответствующими всем известным и проаннотированным на тот момент генам в геноме человека.

Олигонуклеотидная матрица GeneChip использует наборы синтезированных in situ олигонуклеотидных проб, по 11–20 проб в наборе, каждая размером 25 нуклеотидов, для представления транскриптов генов или их изоформ. Для каждого исследуемого гена использованы фрагменты-представители (initial target sequences) длиной 150–450 п.н. для выбора и локализации олигонуклеотидных проб. Уровень экспрессии гена определяется суммой данных всего набора проб (probeset) [154]. Сигнал от пробы с совершенным совпадением всех нуклеотидов учитывается после вычитания неспецифического сигнала кросс-гибридизации от пробы с одним центральным несовпадающим нуклеотидом [155] (см. также http://www.affymetrix.com/support/).

Проблема анализа транскрипции с помощью этого микрочипа в целом связана с рядом технических ограничений и ошибок при создании технологии. Дизайн проб (исходный выбор производителем микрочипов локализации в гене и структуры олигонуклеотидных проб) может не соответствовать целевому транскрипту (гену) и содержать ряд технических проблем, связанных как с гибридизацией, так и с аннотацией – неверное указание гена-мишени, неоднозначность соответствия один набор проб–один ген. Такой дизайн олигонуклеотидных проб может влиять на регистрацию сигналов гибридизации, нормализацию данных, снижать воспроизводимость экспериментов, вести к противоречивым результатам анализа одних и тех же данных [151, 156-159].

Ранее была выполнена независимая аннотация наборов проб микрочипов Affymetrix на основе картирования нуклеотидных последовательностей проб на референсные последовательности генома человека [150, 151, 156]. Выявлен ряд несоответствий в аннотации наборов проб для идентификации генов; такие несоответствия могут затрагивать до 30–50 % наборов проб [151, 157, 159].

Соревнование в сфере технологий производства микрочипов, технологий измерения сигналов экспрессии генов дало большой толчок научным исследованиям и

огромный фактический материал. Отметим еще раз, что за последние годы на смену микрочипам приходят все более совершенные технологии полного секвенирования транскриптом, имеющие ряд принципиальных научных преимуществ, в частности, по способности определения новых вариантов транскриптов гена, по динамической шкале измерения уровня транскрипции [4, 14]. Тем не менее, микрочиповая технология позволяет достаточно надежно и относительно недорого определять дифференциально экспрессирующиеся гены за счет репликации экспериментов [14], и требует разработки специализированных компьютерных инструментов.

В связи с представленными проблемами измерения экспрессии генов встает задача получения статистических оценок качества наборов проб микрочипов, в частности платформы Affymetrix U133 Plus 2.0 для генов человека. Компьютерная оценка уникальности нуклеотидных последовательностей проб микрочипа и соответствия аннотации генов в геноме должна быть подкреплена анализом экспрессии на клинических экспрессионных данных. Такие данные должны быть систематизированы и представлены в общедоступной компьютерной базе данных вместе со статистической оценкой величины и качества измеряемых сигналов на микрочипе, что и было выполнено в настоящей работе.

Интерес представляет исследование расположения проб в геноме, перекрывающихся с цис-антисенс транскриптами генов человека и описание таких транскриптов в целом. Феномен цис-антисенс транскрипции в геноме человека должен быть подкреплен микрочиповыми данными. Сходство нуклеотидных проб с транскриптами повторяющихся последовательностей SINE и LINE в геноме позволяет оценить как качество проб микрочипа, так и возможную экспрессию транспозонов, детектируемую на микрочипе в опухолевых клетках. По методическим и техническим причинам геномные повторы обычно исключаются из дизайна микрочипов [160], в частности из-за избыточности мобильных элементов в геноме и сложности подбора уникальных проб. Таким образом, их потенциальная транскрипционная активность остается недостаточно охарактеризованной, несмотря на многочисленные наблюдения присутствия транскрипции в различных тканях при заболеваниях человека [161], в том числе при раке молочной железы [162].

1.3 РЕГУЛЯТОРНЫЕ УЧАСТКИ ГЕНОВ: ПРОМОТОРЫ И ЭНХАНСЕРЫ

1.3.1. Промоторы и энхансеры

Промотором называют последовательность ДНК, связывающую РНК-полимеразу и служащую отправной точкой транскрипции [163]. В целом, у многоклеточных эукариот, в пределах 100-200 п.н. перед стартом транскрипции выявлена сложная схема промоторных элементов, представленных короткими нуклеотидными последовательностями - мотивами или боксами [164]. Минимальный промотор (или «коровый» промотор, от англ. core) содержит ряд коротких функционально значимых последовательностей размером 5-25 п.о. [126, 165]. В коровом промоторе наиболее полно изучены TATA-бокс, инициатор (Inr-элемент), CAAT-бокс и GC-бокс [166] (рис. 1.5). TATA-бокс представляет собой A/T-богатую последовательность (TATAWAW) [167], расположенную на расстоянии 28-34 п.о. выше старта транскрипции. Inr-элемент содержит старт транскрипции. Выделяют также полипиримидиновый инициатор (TCT), BRE элемент (TFII-B Recognition Element), MTE элемент (Motif Ten Element), DPE элемент (Downstream Promoter Element) [168] и E-бокс. Отметим, что в промоторах эукариот в целом нет однозначной записи контекстных регуляторных сигналов и нет заранее заданной локализации этих сигналов [169], что ставит целую серию задач компьютерного поиска и распознавания таких регуляторных участков в геноме.

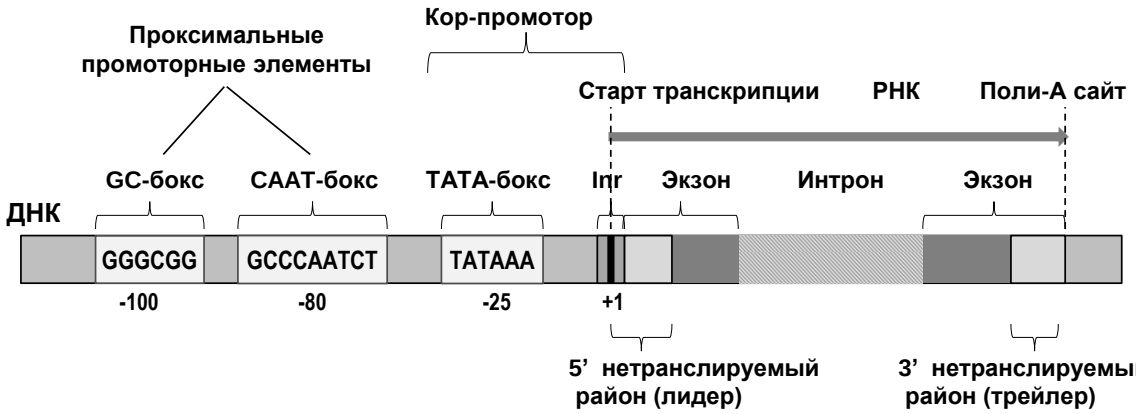


Рис. 1.5. Структура промотора гена эукариот и основные промоторные элементы - GC-бокс, CAAT-бокс, TATA-бокс, Inr. Адаптировано из [170].

Традиционно, по наличию или отсутствию TATA-бокса промоторы делятся на две группы: TATA-содержащие и TATA-несодержащие [166, 171]. В отдельную группу выделяют промоторы, содержащие DPE элемент, являющийся функциональным аналогом TATA-бокса, который локализован в районе +30 относительно старта транскрипции [172].

Транскрипция генов, считываемых РНК-полимеразой III, определяется промотором, лежащим внутри гена. РНК-полимераза III способна к реинициации, транскрибирует гены в районах свободных от нуклеосом [173]. Промотор для РНК-полимеразы I, транскрибирующей гены рибосомальных РНК, недостаточно охарактеризован, не удается составить для него общую схему регуляторных элементов [174].

Энхансер (от англ. enhancer) - это регуляторная последовательность нуклеотидов, усиливающая активность промоторов эукариот [163]. Энхансер значительно, в десятки раз, усиливает транскрипцию, причем это воздействие практически не зависит от расположения энхансера относительно контролируемого гена [175]. Энхансеры способны действовать на больших расстояниях (более нескольких тысяч п.н.), располагаясь как в 5'-, так и в 3'-конце районе, а также внутри гена в составе интронов [176, 177]. Энхансер гена альбумина находится перед промотором, у генов иммуноглобулинов регуляторные элементы расположены в интронах. Энхансер может быть расположен и ниже гена на большом расстоянии, как у бета-глобинового гена [178]. Транскрипция гена *SHH* человека [179] контролируется его энхансером, который расположен на расстоянии 1Мб и вложен в интронный район *LmbR1*. Точечная мутация в этом энхансере вызывает преаксиальную полидактилию, общее врожденное нарушение формирования конечностей у млекопитающих [180].

Потенциальные энхансеры могут быть определены экспериментально с помощью высокопроизводительных экспериментальных подходов [181][182], но нерешенной остается проблема сопоставления энхансеров и их генов-мишеней находящихся на удалении сотен килобаз. Многие дальние энхансеры могут быть вложены в интронные районы других дистально расположенных генов [183], делая неоднозначным соотнесение энхансеров их генам-мишеням.

В литературе обсуждаются два основных механизма действия энхансеров [163]: сверхспирализованные хромосомные петли и непосредственные взаимодействия. Полагают, что функциональные участки генома, содержащие один или несколько генов, образуют длинные петли, включающие десятки тысяч нуклеотидных пар ДНК. Высказано предположение, что такие хромосомные петли закреплены в матриксе клеточного ядра и сверхспирализованы. В состав матрикса входит топоизомераза II, по-видимому, определяющая топологию петли ДНК. Взаимодействие энхансера с белками может менять конформацию всей петли, включая удаленный от энхансера участок ДНК, в результате чего в составе петли изменяется локальная структура хроматина и облегчается транскрипция гена.

Действие энхансера, может приводить к образованию больших транскрипционных комплексов, с которых транскрипция инициируется и ре-инициируется. Такое предположение ставит проблему прямого экспериментального исследования контактирующих участков ДНК, взаимодействующих с комплексом полимеразы II, что было представлено в работе автора на основе технологии секвенирования ChIA-PET [12].

Ремоделирование нуклеосом также может быть одним из механизмов проявления энхансерной активности [163]. В целом можно сказать, что энхансер действует на любой ближайший к нему промотор. Показано, что проксимально расположенные друг к другу гены имеют тенденцию быть совместно регулированными независимо от функциональных различий между ними [184-186] (так называемые «нейтрально ко-экспрессирующиеся кластеры»). Такие кластеры могут появляться в результате эффекта нейтральной коэволюции [187].

Факторы ремоделинга хроматина и белков, организующих структуру хроматина - Ini1, Brg1, CTCF [188] и Rad21 [189] ассоциированы с энхансерными районами. Показано, что Ini1 и Brg1, две субъединицы комплекса SWI/SNF, вовлечены в транскрипционные петли [190, 191].

Эффект регуляции достигается посредством сборки на последовательностях энхансера белкового комплекса, который иногда называют энхансеосомой [192], и его взаимодействия с основным транскрипционным комплексом путем белок-белковых взаимодействий [193]. Обсуждалась проблема стереоспецифичности во взаимодействии энхансерных белковых комплексов - то есть специфичность ориентации в цепи ДНК энхансера по отношению к промотору [175, 193].

Энхансеосома определяется как нуклеопротеиновый комплекс, состоящий из различных наборов сайтов связывания ТФ связанных напрямую или опосредованно с энхансерной ДНК [194, 195]. Прототипом энхансеосомы может служить вирус-индуцируемый энхансер гена интерферона-β (*IFN-β*). Этот энхансер связан субъединицами p50 и p65 NF-κB, ATF-2, IRF-3, IRF-7, c-Jun, и архитектурным транскрипционным фактором HMGA. Атомная модель этого комплекса содержащего восемь этих факторов, связанных с ДНК была реконструирована на основе трех кристаллических структур [196].

Контрольные области генов (LCR) часто содержат множественные энхансерные модули, которые варьируют в размерах от 50 нуклеотидов до 1,5 Кб [197]. Каждый из этих модулей может активировать ген на определенной стадии развития или в определенном типе клеток. Один ген может содержать множество энхансерных

модулей, каждый из которых вносит свой вклад в пространственную и временную регуляцию экспрессии гена. Энхансер в зависимости от белкового фактора может начать вести себя и как негативно действующий регуляторный элемент экспрессии гена - сайленсер (от англ. silencer).

В целом, трудно привести четкие различия между энхансерами и элементами промоторов эукариот. Так, в работе [198] на основе анализа большого набора тканей и клеточных линий человека показана возможность предсказания дистальных энхансеров в геноме на основе контекстных свойств промоторов (присутствия специфичных ССТФ). Регуляторные элементы генов, которые первоначально относили либо к промоторам, либо к энхансерам, обладают рядом общих функциональных характеристик, таких как присутствие сайтов связывания транскрипционных факторов, воздействие на экспрессию близлежащих генов. Возникает задача исследования энхансеров в масштабе генома, определения удаленных от генов регуляторных районов, в том числе с помощью полногеномных экспериментальных методов.

Иерархическая организация регуляторных районов эукариот

Особенность регуляторных районов генов эукариот – их иерархическая организация. Два соседних ССТФ могут представлять композиционный элемент. В этом случае их совместное действие согласовано, то есть его комбинаторный эффект значительно отличается от действия каждого ССТФ в отдельности [199, 200]. 5’-регуляторные районы генов эукариот характеризуются также большим размером, достигающим десятков тысяч п.о. [165], что на порядки больше максимального размера регуляторных районов прокариот.

Считывание с одного гена разных вариантов РНК называется альтернативной транскрипцией. Эта особенность регуляции лежит в основе механизма формирования большого разнообразия первичных транскриптов одного и того же генного локуса и, как следствие этого, разнообразия белков, кодируемых одним и тем же генным локусом. В настоящее время известны примеры первичных транскриптов, в которых сплайсинг может проходить по десяткам альтернативных путей [201, 202]. Так, у человека, более 42% генов имеют альтернативный сплайсинг пре-мРНК. Причем значительная их часть кодирует определенные типы молекул (например, клеточные рецепторы), а также белки, выполняющие системные функции в организме, в частности в иммунной и нервной системах [201].

Транскрипционная активность гена зависит от стадии клеточного цикла, функционального состояния клетки, ткани, органа, стадии индивидуального развития,

действия внешних индукторов [121]. В ядре клетки присутствует набор транскрипционных факторов, которые взаимодействуют с регуляторными элементами конкретного гена. В результате формируется уникальный транскрипционный комплекс, обеспечивающий необходимый уровень транскрипции гена в конкретной клеточной ситуации [121, 123]. Блочно-иерархическая организация регуляторных районов генов эукариот (сайты, промоторы, дистальные регуляторные элементы) обеспечивает возможность гибкой регуляции транскрипции за счет включения/выключения отдельных элементов [124]. Примером сложной организации регуляторного района является удаленный энхансер гена *Pou5f1*, кодирующего ключевой фактор плюрипотентности Oct4, связанный 11 различными транскрипционными факторами [203].

Метилирование ДНК

Метилирование - это ферментативная химическая модификация, добавление метильных групп (CH3) в специфических сайтах белков, ДНК и РНК. Одна из наиболее распространенных форм метилирования представляет собой превращение цитозина в 5-метилцитозин в последовательности нуклеотидов CpG [204]. У человека и большинства млекопитающих ДНК-метилирование - естественная модификация ДНК, и воздействует только на цитозин, стоящий перед гуанином, т.е. метилирование происходит только в CpG-динуклеотидах [205]. Метилирование может предотвращать расщепление ДНК в сайте узнавания фермента рестрикции. Реакция ДНК-метилирования катализируется ферментом ДНК-метилтрансферазой, который осуществляет перенос метильной группы с S-аденозилметионина на цитозин, стоящий перед гуанином.

Метилирование - эпигенетический процесс, не меняющий последовательность ДНК [206, 207]. 70-80% всех CpG-динуклеотидов в геноме человека метилированы [205]. Некоторые гены, экспрессирующиеся в эмбриональном периоде, перестают функционировать к моменту рождения; профиль метилирования в тканях может меняться в течение жизни. Большинство CpG-островков (соответствующих промоторам) в норме не метилированы. Метилирование происходит, прежде всего, в районах генома с низкой плотностью CpG динуклеотидов.

Метилированные основания ДНК экспериментально обнаружены еще в 1948 году [208]. Метилированная ДНК высших эукариот содержит в основном 5-метилцитозин [204, 207, 209]. Существует несколько версий о роли метилирования ДНК: контроль экспрессии гена, контроль целостности хромосомы, контроль пре-рекомбинантных

событий, защитный механизм против встраивания в геном чужеродных последовательностей (ретровирусных элементов) [210, 211].

Профиль метилирования, влияющий на функциональное состояние гена, передается в ряду клеточных поколений, в связи с этим развиваются методы определения возраста клеток в ткани - клеточного старения по данным метилирования [212].

Механизм инактивации гена посредством ДНК-метилирования

Механизм инактивации гена посредством ДНК-метилирования связан с функциями белков-метилаз [210]. Белок MeCp2 (methylated-DNA binding protein 2) связывается с метилированной ДНК и включается в комплекс, состоящий из гистоновых белков и деацетилазы [213]. Этот белковый комплекс, в свою очередь, инициирует компактизацию хроматина, что не дает связаться факторам транскрипции с промоторной областью и, следовательно, происходит инактивация гена [211].

Полагают, что метилирование промотора может быть одним из механизмов инактивации генов-супрессоров опухолевого роста в раковых клетках [149]. Список генов, инактивируемых через метилирование промоторной области, включает MyoD, Rb1, VHL, ген p16. Гиперметилирование промоторной области гена - рецептора эстрогенов (ER) обнаруживается в опухолях толстого кишечника [214]. Показано, что аберрантное метилирование промоторного района гена металлопротеиназы-3 (TIM-3) происходит в различных опухолях: раке молочной железы, раке толстой кишки, карциноме почки. В спорадических опухолях молочной железы показана инактивация посредством метилирования гена BRCA1, гена MYOD и гена ER [215, 216]. Таким образом, метилирование промоторной области может являться механизмом инактивации генов-супрессоров опухолевого роста.

В опухолях наиболее надежным методом для оценки частоты метилирования CpG-островков в настоящее время считается метод RLGS (restriction landmark genomic scanning). Для определения метилирования CpG-островов применяют метил-чувствительные рестриктазы (HpaII, HhaI, NotI, SacII, EagI, BssHII) с последующей амплификацией CpG-острова. Современный метод определения геномного метилирования основан на бисульфитной модификации ДНК с последующей метил-специфической амплификацией или секвенированием [212]. Метод основан на том, что бисульфит натрия преобразовывает все неметилированные цитозины в урацил, в то время как метилированные цитозины, стоящие перед гуанином остаются в не модифицированном состоянии.

1.3.2. Компьютерные методы распознавания регуляторных районов генов

Данный раздел обзора литературы представляет компьютерные методы распознавания регуляторных районов генов эукариот. Компьютерный анализ геномных последовательностей дает возможность объяснить особенности структурно-функциональной организации известных районов геномов, позволяет предсказать функциональные сайты во вновь секвенированной геномной ДНК.

Стандарты описания функциональных сайтов

При описании нуклеотидных последовательностей для обозначения классов нуклеотидов используется соответствующая номенклатура – 15-буквенный вырожденный код IUPAC (Таблица П3 в Приложении). Более точным способом представления и анализа выборок выровненных последовательностей длины L являются весовые матрицы размерности L × 4. Элемент f(i,j) весовой матрицы F = |f(i, j)| определяет частоту встречаемости нуклеотида *i* (*i* =1,2,3,4 соответствует символам A, T, G и C) в позиции *j* (*j* = 1,.., L), подсчитанную по выборке выровненных нуклеотидных последовательностей. Оптимизированная весовая матрица *W*= |*w*(i,j)| может быть вычислена в логарифмической форме с учетом ожидаемых частот [166]. Участки последовательностей, сходство которых с весовой матрицей (мотивом) превышает пороговое значение, рассматриваются как потенциальные сайты связывания транскрипционных факторов [217]. Весовые матрицы ССТФ определены в базах данных, таких как TRRD [200], JASPAR [218], TRANSFAC [129], на основе компиляции результатов связывания ДНК с белковыми ТФ в экспериментах с помощью различных технологий.

Существуют другие способы представления оптимального расположения нуклеотидов в сайте для оценки силы связывания [219]. Традиционная весовая матрица (частотная матрица) может быть преобразована в позиционно-специфичная матрицу энергии связывания PSEM (Position Specific Energy Matrix) и обратно, используя экспоненциальную трансформацию [219, 220].

Методы компьютерного распознавания регуляторных районов

Важнейшая задача анализа регуляторных районов – распознавание сайтов связывания транскрипционных факторов и промоторов в геноме по самой последовательности ДНК и по контекстным характеристикам нуклеотидной последовательности. Для обучения программ компьютерного распознавания и определения потенциальных сайтов связывания транскрипционных факторов разрабатываются базы данных регуляторных районов генов эукариот и ССТФ [129,

200]. Проблема распознавания сайтов связывания обусловлена тем, что хотя транскрипционные факторы связываются с ДНК специфично, большинство сайтов имеет лишь небольшую постоянную «коровую» (core) последовательность, составляющую 4-10 п.о., окруженную некоторым числом не постоянно встречающихся нуклеотидов [221]. Показано, что использование зависимостей между нуклеотидами, в частности динуклеотидных матриц позволяет значительно повысить эффективность распознавания сайтов [222].

Несмотря на то, что создано большое количество методов распознавания промоторов РНК-полимеразы II в геномах эукариот [221, 223-225], проблема повышения точности распознавания в целом остается нерешенной. Заметим, что большинство методов было разработано до начала массового секвенирования и обучалось на сравнительно небольших выборках данных.

При рассмотрении связывания ТФ с ДНК in vivo надо учитывать, что многие белковые факторы не работают по отдельности, часто формирую комплексы с другими факторами и таким образом, могут связывать ДНК прямо или опосредованно. В зависимости от архитектуры комплекса транскрипционных факторов, последовательности ДНК, связанные этим комплексом, могут казаться связанными в экспериментах ChIP-chip для каждого ТФ этого комплекса, хотя только один фактор связан с ДНК напрямую. Например, транскрипционные факторы дрожжей Mbp1 и Swi6, формируют MBF комплекс, играющий важную роль в регуляции клеточного цикла [226]. Swi6 связывает Mbp1, а Mbp1 контактирует с ДНК непосредственно, связываясь с последовательностью ACGCGT [227]. Таким образом, важным моментом анализа сайтов связывания является понимание возможности непрямого связывания ТФ с ДНК через другие белки, называемое также “piggy-back”.

При наличии выборок последовательностей ДНК и множественного выравнивания, для распознавания может применяться стандартный поиск консенсусов, матричные подходы и программы, принимающие во внимание специфические структурные характеристики, энергетические параметры цепи ДНК [228]. При отсутствии доступных структур для обучения (множественного выравнивания), могут применяться методы распознавания, не использующие позиционную информацию о нуклеотидах. Это нейронные сети, подходы, основанные на понятии языка сообщений, периодичности распределения динуклеотидов, и другие методы предсказания, не использующие консенсус, например Фурье-анализ [229]. Для первичного поиска мотивов в наборе нуклеотидных последовательностей используются алгоритмы поиска

олигонуклеотидов, частота встречаемости которых повышена по сравнению с ожидаемым по нуклеотидному составу, такие как алгоритм Weeder [230].

Чтобы более полно использовать ограниченные экспериментальные данные по ССТФ, может использоваться метод реализаций [223]. Точность предсказания может быть увеличена за счет учета окружающего контекста, в котором встретился сайт [231] и учета зависимости между нуклеотидами [232]. Было предложено несколько вычислительных подходов для решения проблемы комбинаторной регуляции транскрипции [233], включая компьютерный отбор специфичных олигонуклеотидов [234], исследование ассоциаций между ними [235]. Заметим, что анализ данных о ССТФ в полногеномной шкале приобретает качественно новый характер, и разработанные ранее статистические подходы являются только основой для более сложных моделей предсказания.

Метод скрытых марковских моделей

Различные реализации метода скрытых марковских моделей (Hidden Markov Model) [29] в настоящее время широко применяются для выравнивания последовательностей ДНК и белков, а также для выявления гомологии между ними, поиска и распознавания последовательностей, обладающих обобщённым сходством [236, 237].

Под марковской цепью в общем смысле подразумевается последовательность событий, каждое из которых происходит с определённой вероятностью. В марковской модели нулевого порядка для последовательности ДНК состояния определяются нуклеотидами, каждый следующий символ в последовательности не зависит от предыдущего (зависит от нулевого числа символов). Модель первого порядка определяет зависимость каждого символа X_i в последовательности X длины L только от одного предыдущего ему. Вероятность P(X) наблюдения последовательности X выражается как:

$$P(X)= P(X_L|X_{L-1})P(X_{L-1}|X_{L-2}) \ldots P(X_2|X_1)\, P(X_1) = P(X_1) \prod_{i=2}^{L} a_{Xi-1Xi} \ . \qquad (1.4)$$

Здесь a_{Xi-1Xi} – вероятность получить символ X_i при условии, что предыдущий символ X_{i-1}. В общем случае вероятность перехода выражается как $a_{st}=P(X_i=t|X_{i-1}=s)$.

Однородная скрытая марковская модель, которая может быть использована для распознавания генов и функциональных районов [29, 238], состоит из множества состояний марковской цепи (например кодирующие и некодирующие участки), переходов между ними, а также из множества наблюдаемых символов и

соответствующих условных вероятностей [237]. Термин “скрытость” в названии метода подразумевает, что в анализируемой последовательности наблюдаемыми являются только характеристики отдельных символов, а не информация о том, в каком состоянии они находятся. Каждому символу в последовательности может быть сопоставлено некоторое состояние. Задача решается через определение условной вероятности по частотам. На практике используют логарифм вероятности. Задача оптимального выбора последовательности состояний решается с помощью алгоритма динамического программирования Витерби [239].

Обзор программ распознавания промоторов

К настоящему моменту в мире создано большое количество компьютерных методов распознавания промоторов РНК-полимеразы II в геномах эукариот [225, 240, 241]. В таблицах Приложения приведён список доступных в Интернете программ поиска генов и программ распознавания промоторов, а также других особенностей контекста, связанных со структурой генов (сайтов связывания транскрипционных факторов, сайтов связывания с ядерным матриксом – MAR, CpG-островов). При разработке компьютерных методик распознавания используются алгоритмы распознавания промоторов, учитывающие специфические особенности их организации [241, 242], в том числе конформационные особенности ДНК [243, 244], особенности структуры хроматина. Повысить точность распознавания промоторов может учет специфики нуклеосомной упаковки промоторов, позиционирования нуклеосом [245], доменной упаковки нитей хроматина, связывания ДНК с элементами ядерного матрикса (MAR) [246].

Для распознавания функциональных участков ДНК (разделения сайтов связывания ТФ и контрольных последовательностей) по контекстным характеристикам использовался дискриминантный анализ [247]. Выбор характеристик для распознавания – переменных в дискриминантном анализе – зависит от конкретной реализации метода [248-250]; в качестве многомерной переменной могут использоваться частоты олигонуклеотидов.

1.3.3. Предсказание сайтов связывания нуклеосом

Организация хроматина и регуляция транскрипции

Геномы эукариот отличаются не только большими размерами (10^8-10^{12} п.о.), но и более сложной организацией хромосом, связанной с упаковкой ДНК в структуру хроматина [251]. ДНК эукариот упакована в нуклеопротеиновые структуры хроматина

[73], что позволяет эукариотам иметь огромный размер геномов по сравнению с прокариотами, и усложняет механизмы регуляции генной экспрессии. Базовый уровень упаковки геномной ДНК в хроматине представляет собой нуклеосому, состоящую из 1.65 супервитков двойной спирали ДНК длины около 147 пар оснований (п.о.), накрученных на октамер белков (по две копии гистонов H2A, H2B, H3, H4) [252]. Связывающая соседние нуклеосомы линкерная ДНК имеет длину от 20 до 60-100 п.о., она взаимодействует с гистоном H1 [253].

Важная роль нуклеосом в регуляции транскрипции генов обусловливает задачу компьютерного предсказания нуклеосомных сайтов. Анализ экспериментально определенных сайтов связывания ДНК с гистоновым октамером дает основу для такого статистического исследования [111]. Особый интерес вызывает вопрос о контекстной специфичности нуклеосомной ДНК [254] и ее связи с регуляторными районами [255]. Показано, что сайты позиционирования нуклеосом обладают слабо выраженными контекстными характеристиками. Создан ряд алгоритмов распознавания нуклеосомных сайтов по контекстным характеристикам [50, 245, 256, 257].

Перестройки хроматина при инициации транскрипции генов эукариот имеют большое значение для сборки комплекса инициации транскрипции и регуляции генной экспрессии [252, 258]. В связи с важностью нуклеосомного кода для регуляции генной экспрессии изучался вопрос о контекстной специфичности последовательностей, содержащих сайты связывания нуклеосом. Было высказано предположение, что расположение нуклеосом в геноме может контролироваться нуклеотидной последовательностью [254]. В частности, фазирование динуклеотидов с периодом двойной спирали ДНК (около 10.5 п.о.) создает благоприятные условия для равномерного однонаправленного изгиба двойной спирали ДНК, который происходит при формировании нуклеосомы. Множественное выравнивание сайтов формирования нуклеосом [259] выявило периодичности динуклеотидов в нуклеосомной ДНК, при этом наиболее яркий сигнал периодичности был обнаружен для комплементарной пары AA/TT. В дальнейшем было показано, что периодичное расположение некоторых типов динуклеотидов характерно для нуклеосомной ДНК [260-262]. К сигналам позиционирования нуклеосомной ДНК относят предпочтительное расположение определенных мотивов в специфичных позициях в пределах нуклеосомной ДНК [256, 261]. Для различных организмов эукариот была показана связь между короткими олигонуклеотидами и экспериментально определенным положением нуклеосом [263], в частности негативная корреляция положения нуклеосом с содержанием поли(A)-трактов [262].

Одна из наиболее популярных программ предсказания сайтов формирования нуклеосом была предложена в работе E.Segal и соавторов [260], и дополнена в работе [264]. Предложенный метод [260] основан на анализе неоднородности распределения динуклеотидов в нуклеосомной ДНК. Анализ с помощью динуклеотидной весовой матрицы [218] включает в себя определение типа динуклеотида в каждой позиции и оценку этого динуклеотида с помощью веса, рассчитываемого по его частоте встреч в соответствующей позиции в выравнивании выборки сайтов формирования нуклеосом.

Точность модели распознавания сайтов формирования нуклеосом, построенной только с помощью нуклеосомной ДНК оказалась невысокой [265]. В работах [266, 267] с помощью метода опорных векторов (SVM, Support Vector Machine) были проанализированы контекстные особенности последовательностей, формирующих нуклеосомы («нуклеосомная» и «линкерная» ДНК).

Отметим, что имеющаяся в базах данных информация [255, 268] о последовательностях ДНК, для которых экспериментально подтверждено существование нуклеосомной упаковки (сотни последовательностей) мала по сравнению с имеющимися полногеномными данными прямого секвенирования нуклеосомной ДНК, что требует разработки новых компьютерных методов анализа.

1.3.4. Полногеномные методы определения сайтов связывания транскрипционных факторов ChIP-seq и ChIP-PET

В последнее время интенсивно развиваются методы определения сайтов связывания на основе иммунопреципитации хроматина (Chromatin IP, или ChIP) с помощью последующего секвенирования связанной с белком ДНК (ChIP-seq) и микро чиповых технологий (ChIP-on-chip) [18, 269]. Появляются данные по структуре хроматина в масштабе полного генома, связанные с метилированием гистонов, составляющих структуру нуклеосомы (гистоны H3 и H4) [19, 270], а также данные по доступности ДНК для белкового связывания, включая прямое секвенирование нуклеосомных фрагментов после обработки ультразвуком. Существуют методы выделения белковой фракции, связанной с ДНК, с помощью формальдегида и последующего секвенирования (метод FAIRE - аббревиатура от Formaldehyde-Assisted Isolation of Regulatory Elements) [271], методы определения участков разрезания ДНК с помощью DNAse I [272].

Техника хроматин-иммунопреципитации (ChIP) эффективна для определения прямых генов-мишеней посредством изолирования фрагментов ДНК, связанных с белками [273]. 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 (то есть хроматин-иммунопреципитация на микрочипе) [10, 274]. Такой анализ возможен для достаточно больших наборов проб, включая, например, отдельные хромосомы или все промоторные районы генов, известные в геноме.

Полногеномные исследования с помощью экспериментов иммунопреципитации на микрочипах (ChIP-chip), позволили получить картину полногеномного распределения гистонов в геноме дрожжей *S.cerevisiae* [275-277]. Для дрожжей методом ChIP-chip было определено полногеномное расположение сайтов связывания транскрипционных факторов [278, 279]. Для ТФ REST (RE1-silencing transcription factor) с помощью ChIP-chip были определены сайты связывания в геноме мыши [280].

Метод иммунопреципитации хроматина с последующим секвенированием ChIP-seq состоит в следующем [18]: контактирующие молекулы ДНК и белков в клетке фиксируются с помощью формальдегида, вызывающего образование ковалентных сшивок между ДНК и белками (рис. 1.6).

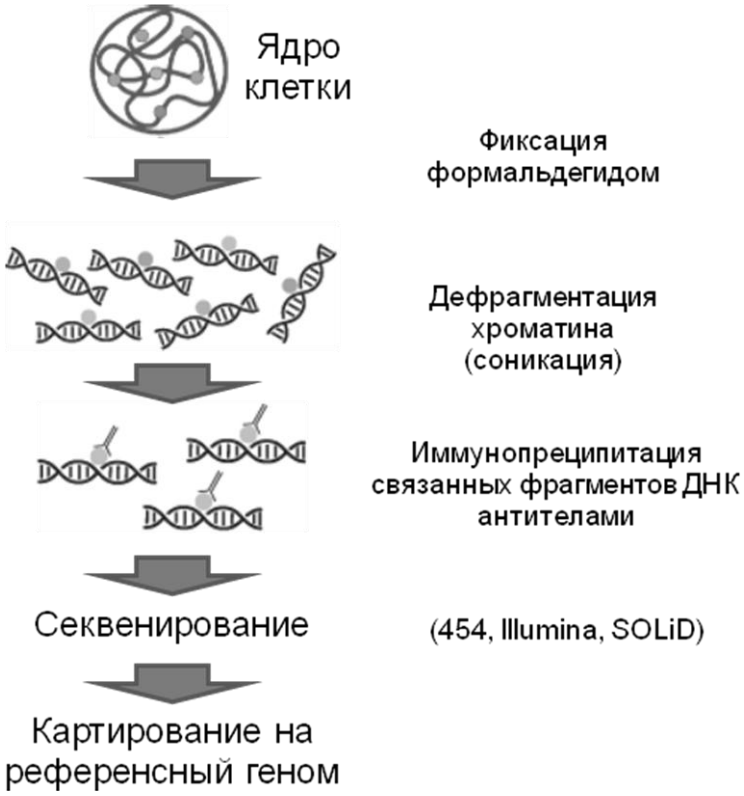


Рис. 1.6. Схема эксперимента ChIP-seq.

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 выделяются фрагменты ДНК, с которыми физически связаны интересующие исследователя белки. Белковая фракция отмывается, а ДНК (относительно короткие фрагменты, не более нескольких сотен оснований) направляется на секвенирование, выполняемое с помощью соответствующего оборудования массового параллельного секвенирования ДНК (Roche 454, Illumina, SOLiD или других современных секвенаторов). Прочитывание ДНК (секвенирование) выполняется не для всей последовательности экстрагированного фрагмента, а для крайних нуклеотидов – от 20 до 50 п.о. Если выполняется лигирование концов, они могут секвенироваться попарно (так называемые парные концы). На следующем рисунке 1.7 представлена схема определения кластеров фрагментов ДНК на хромосоме для парных фрагментов (парных концов метода ChIP-PET).

Картирование секвенированных последовательностей технически требует до нескольких часов машинного времени на персональном компьютере для достаточно больших по размеру геномов эукариот, таких как геном мыши (порядка 3Гб). Для решения этой задачи применяются компьютерные методы оптимизации, быстрого поиска совпадений, существует набор программ картирования от производителей оборудования для форматов данных Illumina [281], и цветовой кодировки SOLiD [269, 282] (см. также обзор программ на SEQanswers, http://seqanswers.com). В частности, вместе с оборудованием SOLiD поставляется пакет BioScope (http://www.solidbioscope.com/). Стандартные программы картирования для форматов Illumina/Solexa, также поставляющиеся вместе с оборудованием, например Genome Analyzer II, включают программы анализа изображений (флюоресценции меченых нуклеотидов на образце по технологии Illumina) Firecrest и Bustard), и программы анализа последовательностей Gerald и Eland. Модуль GERALD (аббревиатура от «Generation of Recursive Analyses Linked by Dependency») предназначен для выравнивания секвенированных последовательностей и фильтрации данных по качеству. Модуль содержит программу ELAND (аббревиатура от «Efficient Large-Scale Alignment of Nucleotide Databases») для картирования прочтений на референсный геном. Существуют альтернативные программы картирования данных Illumina [281], MAQ [283].

Секвенирование парных концов, PET (Paired End Tags) позволяет более точно картировать сайты и имеет ряд принципиальных преимуществ перед односторонним картированием [9, 15] (рис. 1.7).

Рисунок показывает хромосомный профиль связывания транскрипционного фактора p53 в геноме человека, построенный из секвенированных фрагментов ChIP-PET на хромосоме 6 человека [15]. Показано распределение кластеров прочтений на хромосоме, увеличенный участок пика профиля в районе гена CDKN1A, образование «ступеней» из парных фрагментов ДНК, и нуклеотидные последовательности, соответствующие этому пику и содержащие нуклеотидный мотив сайта связывания p53.

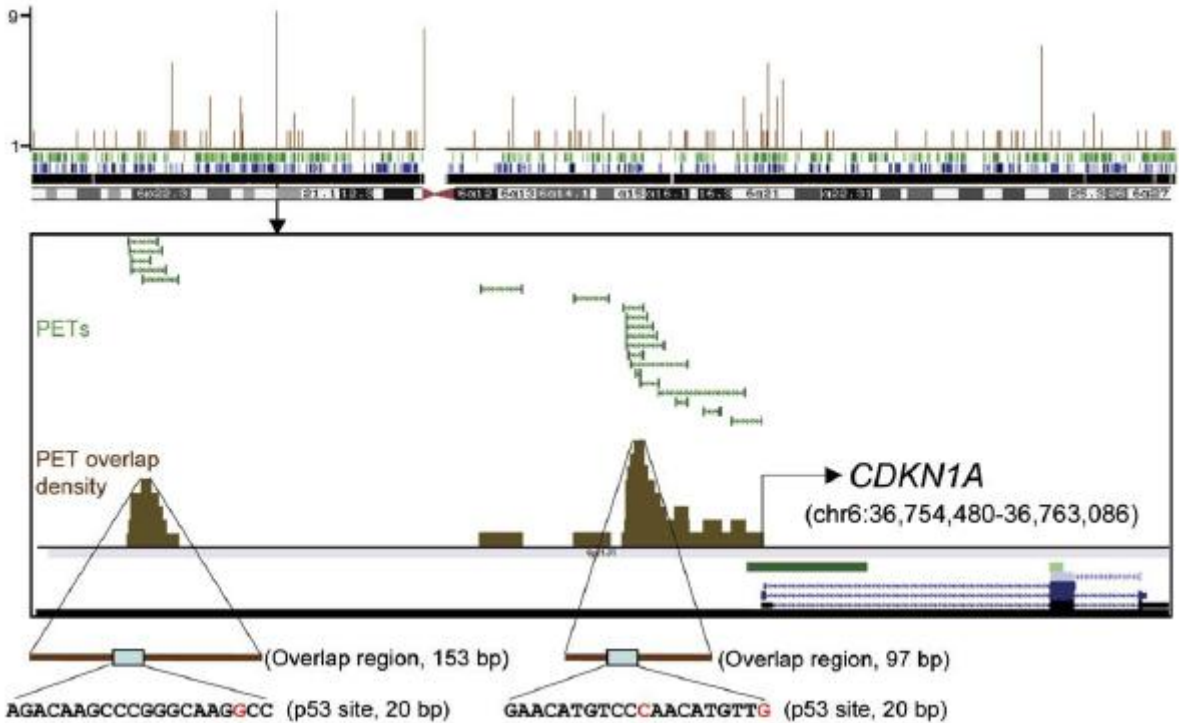


Рис. 1.7. Хромосомный профиль связывания, построенный из секвенированных фрагментов ChIP-PET на хромосоме 6 человека для транскрипционного фактора p53 в геноме человека [15]. Показан увеличенный участок пика профиля в районе гена *CDKN1A*, образование «ступеней» из фрагментов ДНК, и нуклеотидные последовательности, содержащие известный мотив сайта связывания p53.

Относительно недавние работы 2005-2007 годов, использовавшие парные концы, включали клонирование ДНК как необходимый технологический шаг, что замедляло общее время эксперимента и вело к частичной потере данных (при неравномерном клонировании фрагментов ДНК по длине). Технология парных концов с одной стороны имеет преимущества в определении сайтов в тех участках ДНК, где картирование на референсный геном затруднено (например, высоко повторенные, многокопийные участки). В таких случаях парные концы позволяют хотя бы грубо ограничить

локализация сайтов связывания на хромосоме с обеих сторон. В то же время точность локализации сайта не так высока.

ChIP-seq включает прямое секвенирование фрагментов ДНК, связанных с белком, с одной стороны фрагмента. При этом также могут быть парные концы с одновременным картированием двух последовательностей, но исторически аббревиатура ChIP-PET не применяется и относится к работам использовавшим стадию клонирования [9, 15].

Существуют варианты микрочиповых технологий для определения связывания с белками в специализированных вариантах – метод DamID (использующий белок Dam - DNA adenine methyltransferase) [284], метод DIP-seq (DNA ImmunoPrecipitation) для исследования связывания с ДНК без хроматина [285].

Разработан метод GRO-seq (global run-on sequencing) для полногеномного картирования комплексов РНК-полимеразы с учетом ориентации транскрипции [119, 286]. Метод ChIP-exo [120], являющийся модификацией ChIP-seq, и использующий экзонуклеазу для триммирования (разрезания) связанной с антителом последовательности ДНК, позволяет получить высокое разрешение при картировании прочтений и определении сайтов связывания, с точностью до одного нуклеотида.

Метод ChIP-seq появился после ChIP-chip, и обладает рядом преимуществ, основным из которых является принципиальная возможность именно полногеномного анализа – экспериментального определения всех сайтов в геноме через массовое параллельное секвенирование [18]. Таблица 1.2 представляет сравнение методов, использующих иммунопреципитацию хроматина.

Таблица 1.2

Сравнение возможностей полногеномных методов, использующих иммунопреципитацию хроматина

Возможности метода	**ChIP-chip**	**ChIP-PET**	**ChIP-seq**	**ChIA-PET**
Использование проб на микрочипе	+	-	-	-
Парные концы	-	+		+
Использование секвенирования	-	+	+	+
Контакты между удаленными сайтами	-	-	-	+

Секвенирование фрагментов ДНК после иммунопреципитации имеет ряд преимуществ по сравнению с микрочиповыми ChIP-технологиями. Во-первых, результатом ChIP-seq является картина полногеномного распределения сайтов

связывания транскрипционных факторов [18]. Во-вторых, полученный результат свободен от предварительной селекции исходных данных, например от использования только промоторных районов в методе ChIP-chip. В-третьих, результатом ChIP-seq являются не относительные уровни сигнала проб, а позиции сайтов в геномной ДНК, которые могут определены более точно с увеличением глубины секвенирования. После первичной обработки результат ChIP-seq эксперимента оказывается представленным в виде наложенной на геномную последовательность совокупности пиков, соответствующих районам концентрации отдельных коротких фрагментов – прочтений ДНК (или «ридов», от англ. read), полученных в результате единичного акта связывания на пуле клеток.

Уникальность (однозначность) картирования короткой последовательности на геном представляет особую проблему анализа данных. Если в геноме есть повтор (два достаточно длинных участка ДНК в различных локусах), и рассматриваемый короткий фрагмент секвенирования ДНК гомологичен повтору, то однозначное (уникальное) картирование невозможно. Пример таких затруднений – картирование прочтений ДНК для генов, имеющих псевдогены. Возникают термины «картируемость» (mappablity) и «уникома» (uniqueome), последний термин означает часть генома, которая однозначно (уникально) определяется короткими фрагментами заданной длины [87]. Для каждой длины прочтений - своя «уникома» (например, для фрагментов размером 50 нт некартируемых участков гораздо меньше, чем для фрагментов 25 нт). Существуют компьютерные программы и готовые разметки карт «уникальности» для нескольких референсных геномов, в частности генома человека, мыши [87]. Отметим, что программы разметки «уникальности» требуют высокопроизводительных компьютерных вычислений – фактически, поиска всех повторов в геноме.

Для фильтрации последовательностей по качеству разработан ряд программных пакетов [282]. Доступны программы картирования последовательностей в формате SOLiD, позволяющие учитывать ошибки секвенирования: bfast [287] и SHRiMP [288]. Разработаны оценки качества картирования Illumina [289, 290].

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) [283], SOAP (Short Oligonucleotide Alignment) [291], использующие индексы для представления референсной последовательности, отмеченная выше программа ELAND, ориентированная на стандарт данных Illumina.

Распространены также программы Bowtie (используется алгоритм Burrows–Wheeler индекс – BWT [292]), SeqMap [293], RMAP, ZOOM, AMOScmp (amos.sourceforge.net/), программы картирования компании Synamatix (www.synamatix.com), см. также сравнение программ в работе [287].

Есть ряд уже опубликованных программ, например программы – GLITR (GLobal Identifier of Target Regions), MACS [294], SISSRs (Site Identification from Short Sequence Reads) [295], HPeak, PeakFinder, GLITR, QuEST, CisGenome, USeq и PICS (см. для обзора [17]), NEXT-peak [296].

Программа HPeak (Hidden Markov model Peak) использует формализацию скрытых марковских моделей [29] для определения геномных участков, контактирующих с белками [297]. Используется статистическая модель, учитывающая реалистичное распределение вероятностей для прочтений ДНК.

Программа MACS [294] исходно была ориентирована на технологию Illumina Solexa (данные Solexa Genome Analyzer). Программа использует фрагменты (прочтения ДНК) в противоположных ориентациях, чтобы определить так называемый размер сдвига – близость между прочтениями, содержащими сайты связывания. Преимуществом MACS является локальное моделирование «шумового», или контрольного секвенирования с помощью распределения Пуассона по участкам хромосом (участков, размер которых задается пользователем, тысяча – десять тысяч пар нуклеотидов). Отмечено, что программа MACS для проведения процедуры поиска пиков ChIP-Seq, обладает наибольшей точностью в определении сайтов связывания [11].

При определении полногеномного набора сайтов, возможно, что значительная доля этих сайтов нефункциональны и представляют собой следствие биологического шума в передаче регуляторного сигнала [298].

В целом задачи компьютерной обработки полногеномных данных секвенирования, связанных с ChIP-seq, можно разделить на следующие направления [44]:

– 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;

– интегрирование данных секвенирования (кластеры сайтов связывания транскрипционных факторов из различных экспериментов).

С точки зрения организации компьютерных вычислений все большее распространение получают параллельные вычислительные системы. В мире разрабатываются и программы анализа данных ChIP-seq, картирования коротких прочтений, ориентированные на использование параллельных процессоров и программируемых логических матриц FPGA (Field Programmable Gate Array) [299]. Представлены реализации алгоритмов выравнивания коротких прочтений ДНК на GPU - графических ускорителях (graphic processing unit), в частности SOAP3 для описанного выше алгоритма SOAP [300], на графических акселераторах CUDA (compute unified device architecture) [301].

Одна из вычислительных проблем состоит в том, что для многократных компьютерных симуляций положения прочтений в геноме требуются генераторы случайных чисел с очень большим периодом генерации чисел, такие как «Mersenne Twister» [100]. В работе [101] было показано, что «Mersenne Twister» имеет лучшее качество генерации псевдо-случайных чисел при моделировании биологических объектов по сравнению с линейными генераторами, входящими в стандартные программные пакеты.

Выявление и анализ сайтов формирования нуклеосом с помощью массового секвенирования. Бурное развитие методов массового параллельного секвенирования геномной ДНК в последние годы и рост числа полученных с помощью этих методов полногеномных карт расположения нуклеосом дали новый импульс для развития и применения методов предсказания сайтов формирования нуклеосом. Кроме богатого статистического материала для обучения компьютерных методов предсказания, поиска новых закономерностей, связанных с периодичностью, расположением A/T-богатых олигонуклеотидов и подобных методов, возможно прямое выявление позиций нуклеосом в геноме [51, 264].

Представленность нуклеосом отличается между промоторными районами, что связано с присутствием транскрипционных факторов в тех же районах [302, 303]. Развитие новых технологий секвенирования геномной ДНК ставит вопрос о роли самой последовательности ДНК в определении расположения нуклеосом [264]. Zhang и соавторы [304], показали, что контакты гистонов и ДНК не являются определяющими при позиционировании нуклеосом в геноме, а доля контекст-специфичных позиций нуклеосом в геноме составляет 20%. Таким образом, есть расхождение мнений в оценки вклада последовательности ДНК для определения позиций нуклеосом in vivo [257].

Рисунок 1.8 показывает схематическое представление структуры нуклеосомы и периодичности расположения динуклеотидов, которое традиционно используется для компьютерного предсказания нуклеосомных сайтов.

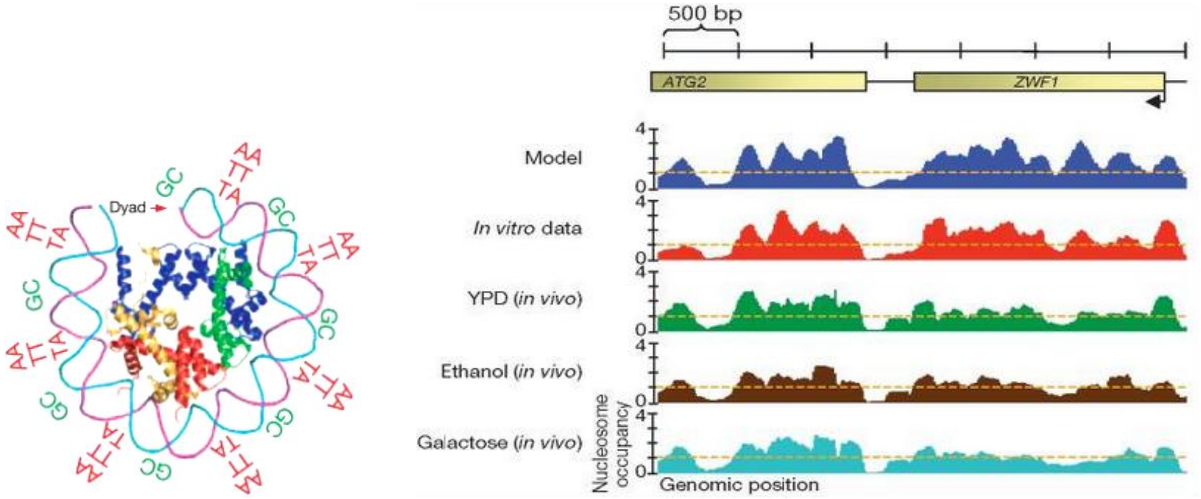


Рис. 1.8. (Левая панель): Схематическое представление структуры нуклеосомы и периодичности расположения динуклеотидов [260].
(Правая панель): Профиль расположения нуклеосом в геноме дрожжей по данным секвенирования. Участок хромосомы 14, район генов *ATG2* и *ZWF1*. Показана модель предсказания (синий) расположения нуклеосом, карта in vitro, и три карты для трех условий роста клеток in vivo (нормальная питательная среда - YPD, этанол и галактоза). Профиль нормирован относительно геномного среднего (y = 1) [264].

На рисунке представлены также опубликованные [264] профили положения нуклеосом, полученные из экспериментов прямого секвенирования нуклеосомной ДНК дрожжей (участок хромосомы 14). Эти профили были повторно рассчитаны в рамках представляемой работы с помощью собственной компьютерной программы.

В недавних исследованиях нуклеосомной ДНК внимание было привлечено к длинным А-богатым районам (политрактам), которые являются ингибиторами (запрещающими сигналами) для связывания с нуклеосомой [264, 276, 305].

Таким образом, необходим анализ новых данных о положении нуклеосом, полученный с помощью современных методов секвенирования, и соответствующих компьютерных программ.

1.3.5. Задачи исследования распределения сайтов связывания транскрипционных факторов в геноме по данным ChIP-seq

Технологии секвенирования ставят ряд новых задач биоинформатики, которые должны быть решены в данной работе. Сюда входит полногеномное определение позиций положения нуклеосом в геноме на основе данных секвенирования ДНК.

Анализ может быть выполнен на примере модельного генома, такого как дрожжи *S.cerevisiae.* Для анализа данных иммунопреципитации хроматина должны

использоваться сайты связывания транскрипционных факторов в промоторных районах генов дрожжей, определенные с помощью технологии ChIP-chip. Положение этих сайтов должно быть проанализировано относительно теоретически предсказанной и определенной в эксперименте нуклеосомной упаковки (позиций нуклеосом в геноме).

Для анализа объемных экспериментальных данных ChIP-seq должны быть разработаны компьютерные методы анализа геномного профиля, поиска сайтов связывания транскрипционных факторов и построение их полногеномных карт.

Необходима компьютерная модель для оценки полноты эксперимента ChIP-seq и алгоритм статистической оценки нижней и верхней границ общего числа ССТФ в геноме на основе компьютерного анализа экспериментально полученного расположения прочтений ChIP-seq в геноме и распределения, смоделированного с помощью компьютерных симуляций. Такой подход даст возможность оценки качества экспериментов ChIP-seq для выявления ССТФ при заданной глубине секвенирования и размере генома для типовых экспериментов на модельных организмах эукариот, таких как мышь, дрожжи.

В целом встает задача разработки компьютерных программ для анализа данных иммунопреципитации хроматина с последующим массовым параллельным секвенированием - экспериментов ChIP-PET и ChIP-seq, и распознавания на этой основе сайтов связывания транскрипционных факторов (ССТФ) в геномах человека, мыши, дрожжей, рыбы *Danio rerio*. Для исследования действия удаленных энхансеров должно быть исследовано распределение сайтов связывания различных транскрипционных факторов в одном и том же типе клеток.

1.4. ТРАНСКРИПЦИОННЫЕ ФАКТОРЫ – ОНКОГЕНЫ И ПРОБЛЕМЫ ИССЛЕДОВАНИЯ ИХ РЕГУЛЯЦИИ

Распределение сайтов связывания транскрипционных факторов - онкогенов

Регуляции экспрессии генов эукариот осуществляется посредством связывания белковых факторов транскрипции с ДНК. Такое связывание может быть как в промоторных районах генов-мишеней, проксимальных к старту транскрипции, так и в удаленных (дистальных) районах. Проблемы определения дистальной регуляции представляют наибольшую сложность, поскольку трудно определить ген-мишень воздействия транскрипционного фактора. Физически связывание белка с ДНК может происходить в удаленных районах, не оказывая влияния на экспрессию генов. Встают

вопросы анализа распределения сайтов связывания транскрипционных факторов в геноме: сколько потенциальных мест связывания может быть, сколько из них реально занято (оккупировано) белком на хромосомах, насколько нуклеотидный контекст влияет на силу связывания.

В данном разделе рассмотрены несколько транскрипционных факторов – онкогенов, исключительно важных для медицинской диагностики, предсказания хода лечения серьезных раковых заболеваний. Исследование распределение сайтов связывания этих транскрипционных факторов в геноме человека представляет актуальную проблему биоинформатики. Рассмотрены транскрипционные факторы p53, c-Myc (MYC), STAT, ERα, для каждого из которых были выполнены эксперименты по определению генов-мишеней, подробно описанные в следующих главах данной работы.

1.4.1. Транскрипционные факторы p53, STAT1, FOXA1

Белок p53, кодируемый геном TP53, является транскрипционным фактором [15, 306, 307]. Молекулы белка p53 образуют тетрамер, способный активировать транскрипцию ряда генов, имеющих соответствующий сайт связывания этого ТФ. Элемент ДНК, с которым связывается p53, состоит из двух расположенных друг за другом на расстоянии от 0 до 13 нуклеотидов "полусайтов", имеющих обобщенную структуру в 15-буквенном алфавите: RRRC(A/T)(A/T)GYYY [308] (см. также рис. 1.7). Общее представление сайта: 5'-RRRCWWGYYY-N(0-13)-RRRCWWGYYY-3'.

Впервые ядерный ДНК-связывающий белок p53 был идентифицирован в составе комплекса с Т-антигеном SV40 в 1979 г. [309]. Ген TP53 человека эволюционно консервативен [310] В дальнейшем было установлено, что p53 экспрессируется на высоком уровне практически во всех типах опухолей, играет существенную роль в широком круге клеточных процессов и является геном-супрессором [311].

Транскрипционная репрессия является функцией С-концевой части молекулы p53 и обусловлена, в том числе, способностью связываться с базальным компонентом транскрипционного аппарата - фактором TBP [312] и подавлением активности TFIID. Кроме того, p53 репрессирует активность нескольких транскрипционных факторов, среди которых Spi1, HIF-1 [313], рецептор тироидного гормона [314], рецептор эстрогенов [315], транскрипционный фактор STAT5, гены BCL2, RELA, MDR1, Hsp70, MAP4 [15].

Транскрипционный фактор STAT1 (signal transducer and activator of transcription protein 1) активируется после связывания цитокинов или интерферонов с их

рецепторами. Связывания цитокина с его распознающим рецептором индуцирует фосфорилирование рецептора Jak киназами [316]. Такие фосфорилированные тирозины обеспечивают сайты докинга для белков семейства STAT. Белки STAT фосфорилируются сами, отделяются от рецептора и могут димеризоваться. В форме димера этот транскрипционный фактор может транслоцироваться в ядро клетки, где он модулирует экспрессию своих генов-мишеней. Примечательна скорость работы системы активации – ДНК-связывающая активность STAT может детектироваться через минуты после связывания цитокинов [317]. Белки STAT обладают шестью существенными функциональными возможностями. Это способности: (1) связывать фосфорилированные тирозины, (2) быть фосфорилированными по тирозинам, (3) димеризации, (4) транслоцироваться в ядро, (5) связывать ДНК, и (6) модулировать экспрессию генов.

Связывание STAT1 с ДНК в геноме человека (клетки HeLa S3) исследовалось с помощью методов иммунопреципитации хроматина ChIP-chip и ChIP-seq [316]. Установлены предпочтения связывания различных белков семейства STAT [318] к сайту TTC(N_x)GAA в зависимости от размера спейсера, составляющего 3 или 4 нуклеотида.

Модель связывания STAT1 послужила для проверки методов анализа пиков и определения сайтов, таких как SISSRs [295] и NEXT-peak [296].

Транскрипционный фактор FOXA1 был предложен в качестве так называемого «первооткрывающего» фактора (pioneering factor) [319, 320] который потенциально может направлять связывание других факторов, в частности ERα. Интересно отметить совпадение мишеней геномного связывания факторов семейства STAT и FOXA1 по данным мета-анализа нескольких ChIP экспериментов [321].

Белок CTCF (CCCTC-binding factor) является инсулятором, т.е. необходим для ограничения транскрипции генов в изолированных областях генома [296]. Показана связь CTCF с доменами хроматина, и привлечением других факторов, в том числе FOXA1 и ERα [322].

1.4.2. Транскрипционный фактор c-Myc

Протоонкоген MYC кодирует транскрипционный фактор c-Myc (здесь и далее называемый Myc), который регулирует размер клеток, клеточную пролиферацию и апоптоз [68, 323, 324]. В норме митогены индуцируют экспрессию Myc, когда клетки входят в клеточный цикл [68], и наоборот клеточное «молчание» (задержка клеточного цикла) и дифференцировка значительно уменьшают экспрессию Myc. Напротив,

опухолевые клетки имеют генетические нарушения регуляции экспрессии гена Myc; постоянная экспрессия Myc является центральным моментом для их трансформации. Myc – это белок класса «лейциновая застежка» (спираль-поворот-спираль), который димеризуется с белком Max, облигатным партнером для активации транскрипции [325]. Мотив связывания с ДНК это 5’-CACGTG-3’, известен также как E-бокс. Myc также подавляет транскрипцию через взаимодействие с Miz-1 или через другие элементы «корового» промотора [326]; заметим, что механизмы подавления репрессии недостаточно изучены.

Помимо облигатного партнера связывания Max, фактор Myc взаимодействует с другими транскрипционными комплексами и транскрипционная активация генов воздействием Myc модулируется через эти взаимодействия [327]. На рисунке 1.9 представлен механизм активации и функциональное воздействие Myc.

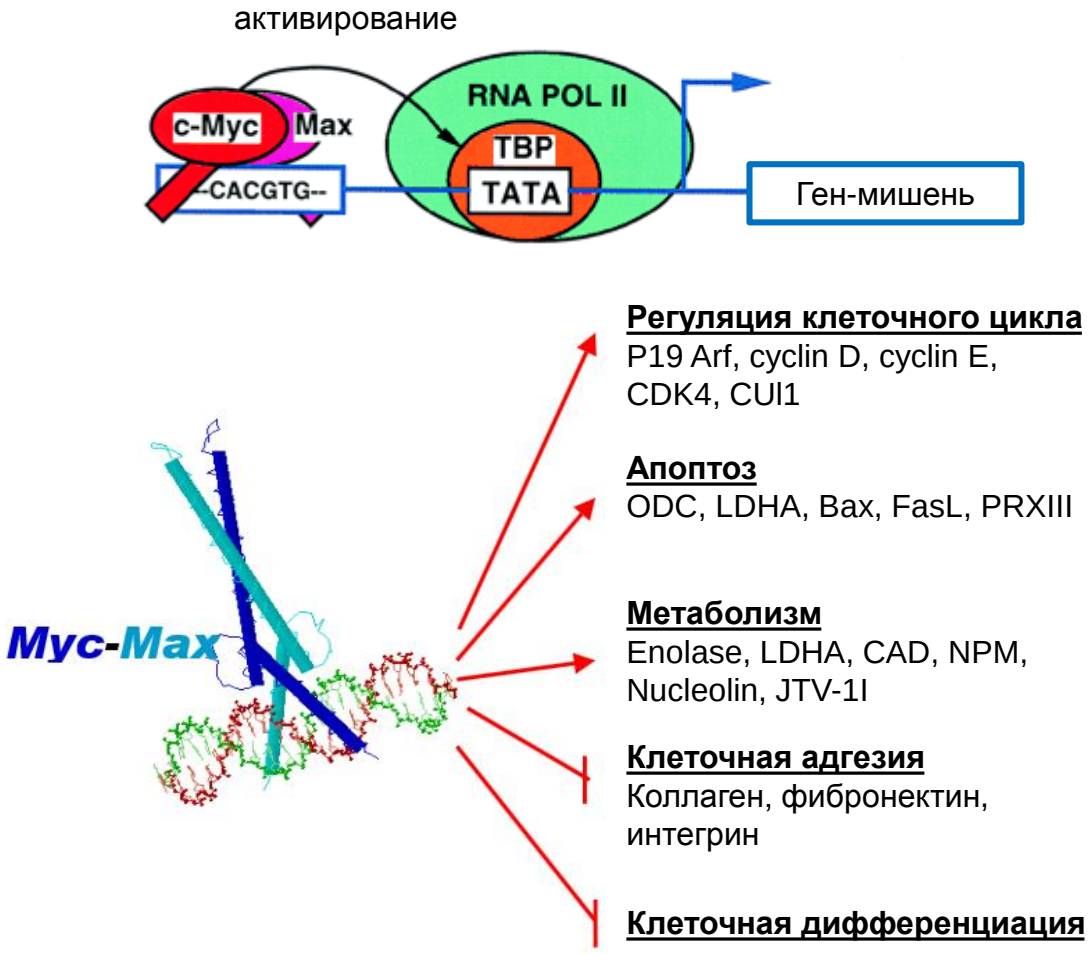


Рис. 1.9. Ген Myc - механизм активации и функциональное воздействие. Гетеродимер Myc-Max класса «лейциновая застежка» связывается с сайтом ДНК E-бокс (CACGTG), регулируя набор генов клеточного цикла, апоптоза и клеточной адгезии. Показаны примеры генов-мишеней с различными клеточными функциями [331].

Транскрипционная активность Myc критична для его способности вызывать опухолевую трансформацию, поскольку аллели с транскрипционно не

функциональным MYC имеют значительно уменьшенный потенциал трансформации [328]. Ряд работ был посвящен исследованию роли Myc в образовании опухолей и развитии, идентификации генов-мишеней воздействия Myc, изучению того, как транскрипционные изменения этих мишеней ведут к увеличению размера клеток, прогрессии клеточного цикла, апоптозу, нарушению дифференцировки клеток [329]. Известно около 1500 генов, показанных как отвечающие на воздействие Myc, и представленных в базе данных его генов-мишеней (www.myccancergene.org) [330]. Высокопроизводительное определение профилей экспрессии генов в клетке, включая микрочипы, и метод SAGE (Serial Analysis of Gene Expression - серийный анализ экспрессии генов) применялись для определения генов ответа на Myc. Поскольку большинство исследований ограничено возможностями тестирования методом количественной ПЦР (qPCR) и неспособны точно различить прямые и непрямые гены-мишени, только небольшая часть генов ответа на Myc считается его прямыми генами-мишенями.

Протоонкоген MYC кодирует онкогенный транскрипционный фактор, играющий центральную роль в образовании многих видов рака. MYC принадлежит к семейству генов myc, которое включает также Bmyc, MYCL и MYCN. Впервые идентифицированный как клеточный гомолог ретровирусного онкогена v-myc, ген MYC, связан с хромосомными транслокациями и амплификации в клетках опухолей человека.

ДНК-чипы, также как и другие методы определения дифференциальной экспрессии генов, привели к накоплению данных о генах ответа на Myc. Эти гены могут кластеризоваться в различные группы транскриптов, которые клеточно- и видоспецифичны. Требуется дальнейший анализ, регулируется ли их экспрессия напрямую или через посредников. Такие данные были собраны в базе данных генов ответа на Myc (Myc Target Gene database) [330]. Отмечается, что в этой базе данных гены-мишени c-Myc были установлены с помощью одного или нескольких экспериментов определения дифференциальной экспрессии генов, включая технологии SAGE [332], ДНК-чипы [333]. Большинство из предполагаемых генов-мишеней, собранных в этой базе данных, не имеют дополнительных подтверждений, являются ли они прямыми мишенями c-Myc или мишенями, опосредованными другими транскрипционными факторами.

При объединении методов иммунопреципитации хроматина с методом детекции на микроматрицах (ChIP-chip) или ChIP-qPCR (qPCR продуктов иммунопреципитации), могут быть определены локусы прямого связывания Myc в геномах [334, 335]. Однако

такие методы использования ChIP сфокусированы только на некоторых специфических характеристиках и участках генома. По совокупности геномных исследований ранее полагали, что Myc может регулировать до 10–15% всех генов человека.

Научной проблемой было определение прямых мишеней и, соответственно, клеточных метаболических путей, на которые влияет Myc в конкретной экспериментальной системе. Были предложены различные стратегии для получения дополнительной информации о прямой регуляции Myc для небольшого числа генов-мишеней. Они включают регуляцию химерным белком Myc-эстрадиоловый рецептор (MycER) в присутствии или отсутствии циклогексимида, тестирование промоторного репортерного гена, экспрессия с последующей стимуляцией сыворотки и корреляцией экспрессии исследуемого гена с экспрессией Myc в различных клеточных системах [330]. Тем не менее, такие подходы не дают определенного доказательства, является ли ген прямой транскрипционной мишенью Myc. Только метод иммунопреципитации хроматина (ChIP) дает идентификацию геномных последовательностей, связанных с Myc в эксперименте in vivo, и приводит доказательство того, что ген напрямую регулируется Myc [336].

1.4.3. Транскрипционный фактор - рецептор эстрогенов

Определение генов-мишеней рецептора эстрогенов имеет огромное значение в онкологии. Рецептор эстрогенов связан с раком молочной железы (раком груди). Рецептор представляет собой транскрипционный фактор, индуцируемый появлением в среде (или изменением концентрации) соответствующего гормона эстрогенов - эстрона (E1), эстрадиола (17-β-эстрадиола (E2) у человека) и эстриола (E3). Для проявления полной транскрипционной активности необходимо фосфорилирование рецептора эстрогенов [337]. Рецептор эстрогенов, как и другие транскрипционные факторы - рецепторы стероидов, взаимодействует со структурными компонентами хроматина и "рекрутирует" (привлекает) другие ядерные белки - эстроген-индуцируемые факторы транскрипции [338, 339].

Присоединение гормона к рецептору вызывает диссоциацию комплекса рецептора с белком HSP90. Наиболее важны конформационные изменения гормон-связывающего участка, определяющие специфичность реакции в тканях организма, различные при взаимодействии с агонистом - эстрогеном или с антагонистом [340]. Далее рецептор в виде гетеродимера взаимодействует с ERE - элементом ответа на эстроген (эстроген-чувствительным элементом) (ERE - estrogen response element) [320].

Ген ESR1 человека, кодирующий рецептор эстрогенов альфа - ERα, расположен на 6 хромосоме, содержит 8 экзонов. Рецептор эстрогенов стандартно подразделяется на несколько доменов, различающихся по функциям [341] - A/B, C (лиганд-связывающий домен), D и E/F (лиганд-связывающий домен) (см. рис. 1.10).

Область A/B содержит структуру, обладающую трансактивационной или активационной функцией (AFI), мало зависящей или не зависящей от лиганда. Следующий участок (C) обеспечивает связывание гормон-рецепторного комплекса с ДНК и димеризацию рецепторов. Связывание белка с ДНК реализуется через специфические последовательности ДНК, эстроген-чувствительные элементы (ERE). Для связывания нужны ионы цинка [338]. Далее в структуре белка расположен домен D, а после него находится многофункциональная область E, содержащая гормон-связывающий участок и лиганд-зависимую структуру с трансактивационной/активационной функцией (AFII), затем расположен домен F, поддерживающий функцию AFII.

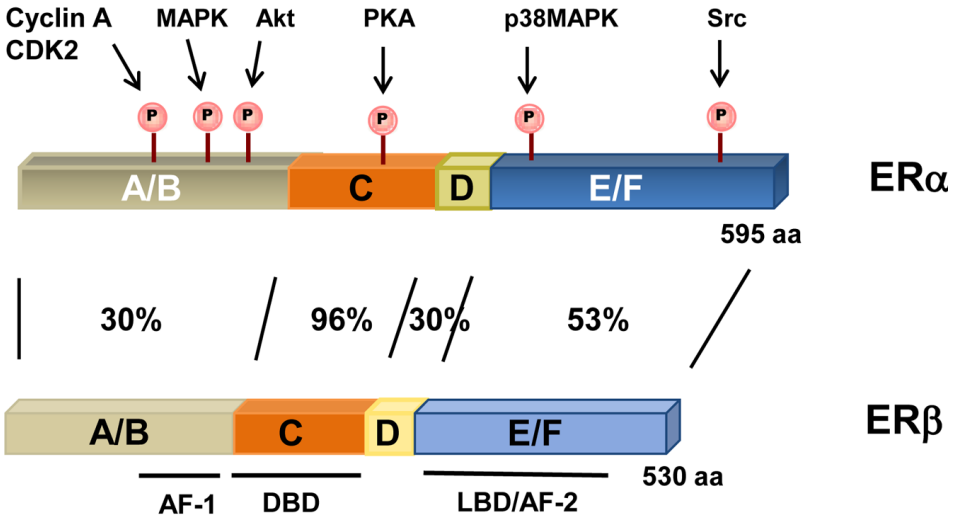


Рис. 1.10. Структура доменов транскрипционных факторов ERα и ERβ [341]. Оба белка имеют четыре функциональных домена, включая ДНК-связывающий домен (DBD), лиганд-связывающий домен (LBD) и два участка с функцией транскрипционной активации (AF-1 и AF-2). Показан процент гомологии этих доменов между ERα и ERβ и локализация нескольких сайтов фосфорилирования в ERα. Отмечены киназы, модулирующие действие рецептора: Akt, серин/треонин специфичное семейство протеин-киназ; CDK2, циклин-зависимая киназа 2; MAPK (mitogen activated protein kinase), активируемая митогеном протеин-киназа; PKA, протеин-киназа A; Src: ко-активатор стероидного рецептора.

Для эстроген-зависимых генов стабилизация связывания ERα с ДНК обеспечивается участием факторов, связывающихся с дистальными энхансерными сайтами. Стероид-зависимые энхансеры активируются, будучи связанными с гормон-рецепторными комплексами (особенно при димеризации), а для перевода в активное состояние стероид-независимых энхансеров необходимо взаимодействие с другими

трансактивационными факторами [342]. Трансактивационные домены у «классического» рецептора эстрогенов ERα и у рецептора второго типа ERβ (см. рис. 1.12) не полностью гомологичны, что свидетельствует об их способности активировать различные эстроген-зависимые гены [343]. Сайты ERE с низкой аффинностью больше нуждаются в других транскрипционных факторах для связывания белка с ДНК, что может выражаться в прямых белок-белковых взаимодействиях, как было предложено для AP-1 и ERα [344].

При новообразованиях молочной железы концентрация эстрогеновых рецепторов часто имеет более высокий уровень, чем в нормальной ткани, что показано с помощью микрочиповых технологий Affymetrix для различных гистологических классов опухоли [152]. Еще в 70-х годах высказывалось предположение, что этапу, предшествующему клиническому выявлению опухоли, свойственно усиление экспрессии эутопических (нормальных) рецепторов, на смену чему приходит появление эктопических (измененных) рецепторных белков [345].

Посредством технологии GRO-seq показано, что передача сигнала эстрогена напрямую регулирует значительную часть транскриптома в клеточных линиях опухолей молочной железы, включая активность всех трех типов РНК-полимераз и практически все классы некодирующих РНК [286].

С помощью различных технологий иммунопреципитации хроматина были определены сайты связывания ERα в геноме человека (в основном на клеточной линии аденокарциномы MCF-7): ChIP-PET [346], ChIP-on-chip [339] и [320], ChIP-seq [347, 348]. Встает вопрос сравнения данных этих экспериментов, исследования наиболее полного набора сайтов связывания в геноме человека с помощью ChIP-seq и построения компьютерной модели распределения этих сайтов.

1.4.4. Возникновение опухолей и регуляция транскрипции

Вопрос происхождения опухолей, перерождения клеток и возникновения опухоли из мутировавших клеток коммитированного предшественника или от стволовой клетки имеет принципиальное значение не только для понимания патогенеза, но и для их лечения [349]. Прямое увеличение концентрации химиотерапевтических препаратов или дозы облучения ограничивается чувствительностью стволовых клеток, в первую очередь, костного мозга и кишечника [349, 350]. Встает задача поиска ингибиторов специфического действия онкогенов или интегрированных протоонкогенов для подавления роста опухолей [350]. Злокачественная трансформация возникает не только в результате генетических изменений, преобразующих протоонкогены в онкогены, но и

при потере контроля генов - супрессоров опухолей, присутствующих в нормальных хромосомах. Нормальные аллели генов-супрессоров имеют доминантный эффект и препятствуют появлению трансформаций. Повреждение генов-супрессоров, таких как p53, в результате мутаций приводит к возникновению или прогрессирующему росту опухолей [351].

Существует несколько важных для роста рака молочной железы онкогенов. Рак молочной железы клинически классифицирован в подгруппы, основанные на наличии определенных белков, включая рецептор эстрогена (ER), рецептор прогестерона (PR) и человеческий эпидермальный фактор роста 2 (HER2). Показано, что повышенная экспрессия FAM83B [352] связана с более агрессивной, тройной отрицательной подгруппой рака молочной железы с недостатком ER, PR и HER2. В связи с этим встают задачи, как описания экспрессии этих генов, так и полногеномного анализа их генов-мишеней.

Метилирование ДНК, модификации гистонов и прогрессия опухоли

Метилирование ДНК может менять характеристики нуклеосомной структуры хроматина, зависящей в значительной степени от функционирования комплексов "ремоделинга" [353, 354]. Комплексы ремоделинга могут удалять нуклеосомы, менять их расположение на ДНК, регулировать в нуклеосоме присутствие различающихся по аминокислотной последовательности вариантов гистонов, например H2AZ.

Все три процесса - метилирование ДНК, модификация гистонов и ремоделинг нуклеосом - связаны между собой. Мутации в генах, кодирующих белки, контролирующие эти процессы, могут быть причиной развития разных типов рака.

Метилирование цитозина ДНК с образованием 5-метилцитозина осуществляется в динуклеотидных последовательностях CpG [355]. Метилирование может приводить к подавлению транскрипции из-за связывания белков, узнающих метилированные CpG. Такие белки репрессируют гены в том числе путем связывания гистондеацетилаз, удаляющих ацетильные группы остатков лизина - характерного маркера активного хроматина [356, 357].

Ландшафт метилирования ДНК ("метилома") подвержен изменениям при раке, причем характер этих изменений специфичен для разных видов опухолей [358, 359]. Имеет место гипометилирование (общее «глобальное» снижение метилирования) [360, 361], но для некоторых CpG островков наблюдается локальное избыточное метилирование, характерное для развития рака и приводящее к репрессии генов [359], и к подавлению активности генов-супрессоров опухолей. Глобальное снижение метилирования, показано для глиобластомы человека [362], рака толстой кишки [363].

Модификация гистонов хроматина и прогрессия опухоли

Процессы метилирования ДНК тесно связаны с модификацией гистонов хроматина [364]. Механизм зависимости модификации гистонов от метилирования ДНК основан на образовании комплексов между белками, формирующими характеристики эпигенома: ДНК-метилтрансферазами; белками, узнающими метилированные сайты; гистондеацетилазами и гистонметилтрансферазами [365]. Обнаружена способность ДНК-метилтрансферазы связываться не только с гистонметилтрансферазой, но и с гистондеацетилазой и гетерохроматиновым белком HP1 [366]. ДНК-метилтрансфераза привлекается в состав комплекса белков группы Polycomb (PcG), осуществляющих репрессию генов дифференцировки [367].

Фактор p300 - это гистон ацетилтрансфераза, часто присутствующая в энхансерных районах [368-370]. p300 рекрутируется (взаимодействует с) энхансерами, и это рекрутирование поддерживается Oct4, Sox2 и Nanog [371]. Другой регулятор хроматина, Suz12, относится к группе белков Polycomb, который вместе с EZH2 и EED формирует репрессивный комплекс хроматина PRC2/3 обладающий метилазной активностью для гистона H3 (метилирование лизина в позициях 27 и 9). Suz12 необходим для развития эмбрионов мыши [372].

Как правило, активный хроматин характеризуется ацетилированием остатков лизина (обозначение аминокислотного остатка - K) в гистонах и триметилированием лизина в положении 4 гистона H3 (H3K4me3) [373], тогда как модификации H3K9me3 и H3K27me3 свойственны репрессированному гетерохроматину и неактивному эухроматину, соответственно [270].

Сайленсинг генов, обусловленный характерными модификациями H3K27me3 и H3K9me3, может сопровождаться деацетилированием гистонов и метилированием ДНК. Модификация H3K9me3 привлекает структурный белок HP1, играющий важную роль в установлении сайленсинга генов и образовании гетерохроматиновых структур [365]. Возникновение и метастазирование рака молочной железы ассоциировано со сниженным уровнем экспрессии HP1, связывающего метилированный лизин гистона H3 в позиции 9 (H3K9me3) [374-376].

Общий "рисунок" расположения модификаций в гистонах нуклеосом определяет специфичность взаимодействия с хроматином активирующих или репрессорных белковых комплексов. Последние исследования включают разработку каталога предсказательных «сигнатур» - специфичных наборов модификаций гистонов, характеризующих различные типы рака [377].

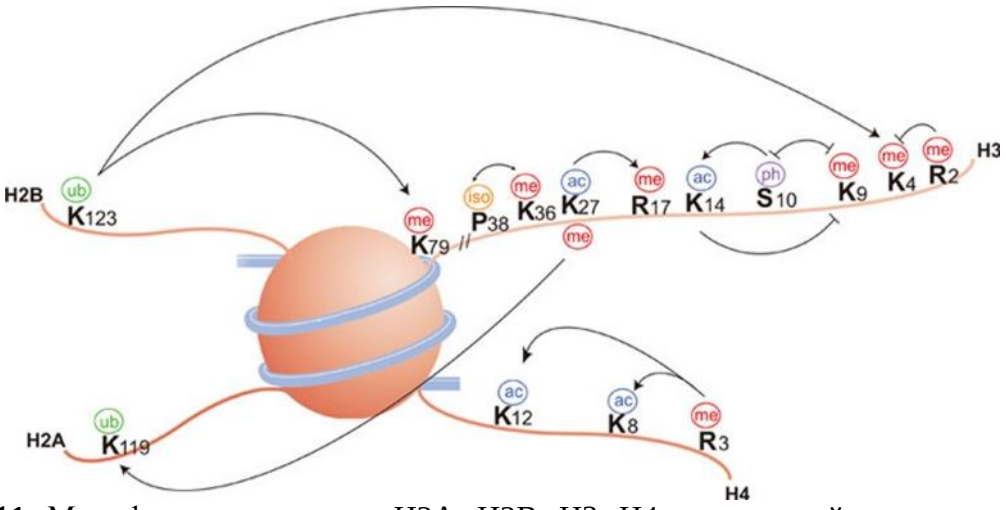


Рис. 1.11. Модификации гистонов H2A, H2B, H3, H4 и взаимодействия между ними [378].

Отметим некоторые исследованные модификации гистонов, связанные с раком. Характерным изменением в модификации гистонов в раковых опухолях человека является потеря ацетилирования K16 и триметилирования K20 в гистоне H4 [20, 379]. Снижение активности гистонметилтрансфераз в отношении сайтов H3K27 и H3K9 коррелирует с нарушениями функционирования белка Rb и с аномалиями регуляции клеточного цикла, наблюдаемыми в опухолевых клетках [365]. При раке наблюдается повышение экспрессии белков комплексов группы Polycomb (PcG), ответственных за модификации гистонов H3K27me3 и H3K9me3 [380]. В норме белки PcG в эмбриональных стволовых клетках необходимы для поддержания состояния плюрипотентности клеток, через репрессию (временное и обратимое подавление) экспрессии генов, начинающих экспрессию при дифференцировке клеток (так называемые бивалентные кластеры генов). В раковых клетках отмечено усиление метилирования промоторов генов-мишеней белков PcG, что препятствует дифференцировке и поддерживает устойчивое самообновление клеток, ведущее к малигнизации [381].

В целом эпигеномные модификации структуры хроматина играют важную роль в возникновении опухолей [380]. Гетерогенность опухолевых клеток обусловлена в значительной степени вариациями эпигеномной структуры хроматина (модификаций гистонов) в клетках-предшественниках опухолей [382]. Отмечалось, что прогрессия опухоли определяется изменениями эпигеномных характеристик, возникающих чаще, чем мутации нуклеотидов [383].

При возникновении рака происходят не только модификации хроматина, но и активизируются мобильные элементы генома, кодирующие РНК [384]. Рассматривается потенциальное влияние на прогрессию опухолей ретротранспозонов семейства L1, не

обладающих длинными концевыми повторами (элементы класса LINE, включая Alu повторы). Элементы LINE составляют 17% генома человека, показано, что до 100 копий потенциально способны к автономным перемещениям [385]. Элементы L1 могут перемещаться в геноме зародышевой линии, во время раннего развития и в некоторых соматических клетках [386]. Обратная транскриптаза и эндонуклеаза, кодируемые элементами L1, могут обеспечивать перемещения неавтономных ретроэлементов Alu (приблизительно миллион копий у человека). Наличие в клетках обратной транскриптазы, кодируемой как этими элементами, так и эндогенными ретровирусами, рассматривается как функциональная характеристика опухолевых клеток [387]. Активация транскрипции L1 связана с инициацией и прогрессией опухолей [388-391]. Активность транскрипции ретровирусов детектируется при раке молочной железы. Была показана связь инициации транскрипции с присутствием в 5′-области генов ретротранспозонов у мыши и у человека [392].

Экспонирование клеток к повреждающим воздействиям ДНК, таким, как химиотерапия или радиация, может вести к индукции транскрипции SINE-элементов, что подтверждает глобальную активацию транспозонов в геноме при стрессе [393, 394].

Терапевтические подходы к воздействию на опухолевые клетки

Поиск участков генома, содержащих вставки ретротранспозонов, в раковых клетках позволяет выделить онкогены и маркеры рака [395]. Показано, что подавление обратной транскриптазы воздействием на РНК снижало пролиферацию и способствовало дифференцировке для некоторых клеточных линий рака [396].

Исследование механизмов миграции и инвазии опухолевых клеток позволяет найти подходы к терапевтическому воздействию на эти процессы, связанные с воздействием на регуляцию активности белков, как на транскрипционном уровне, так и на уровне модификаций белков. Один из таких подходов состоит в воздействии на киназу фокальных контактов (FAK, Focal adhesion kinase) в опухолевых клетках [397-399]. Предложен ряд ингибиторов FAK, блокирующих прогрессию опухоли [399-401]. Поскольку рак как заболевание влечет нарушения паттернов экспрессии генов, то перспективным для терапии является воздействие на транскрипционные факторы - онкогены с избыточной активностью [402]. Прямое ингибирование экспрессии такого транскрипционного фактора (например, через интерференцию РНК), также как нарушение его связывания с ДНК [403] дает антиопухолевый ответ с минимальными побочными эффектами. Исследуются также подходы к получению молекулярных антагонистов интегринов, которые ингибируют пути передачи сигнала этих белков [404], или же стимулирующие развитие сосудов в опухолях [405].

1.4.5. Задачи анализа регуляции транскрипции онкогенов

Для определения генов мишеней действия онкогенов MYC и ERα, а в последующем и мишеней для терапевтического воздействия, необходим компьютерный анализ данных экспериментов ChIP-PET и ChIP-seq определения сайтов связывания транскрипционных факторов MYC и ERα. Работа должна быть выполнена на культурах клеток опухолей человека P493 и MCF-7, как проверенных моделей для индуцируемой экспрессии этих генов, соответственно.

Необходимым элементом является проверка качества определения сайтов в экспериментах ChIP-seq с помощью выборочного тестирования сайтов посредством кПЦР и оценка корреляция между силой связывания транскрипционных факторов ERα и MYC, измеренной с помощью кПЦР и числом прочтений ДНК.

Важен анализ, как мотивов связывания этих факторов, так и сопутствующих нуклеотидных мотивов транскрипционных факторов, связывающихся в окрестностях сайтов ERα. Проблема ассоциации связывания ТФ в геноме должна быть исследована на присутствие маркеров открытого хроматина (маркеров модификаций гистонов, в частности гистона H3 - H3K4me3, H3K4me1, H3K9ac, H3K14ac), определенных с помощью технологии ChIP-seq. Открытое состояние хроматина может быть оценено по отсутствию нуклеосомной упаковки, определяемой с помощью метода FAIRE [271]. Должна быть исследована возможность предсказания сайтов связывания транскрипционного фактора ERα в геноме человека.

Дальнейший анализ такого модельного фактора, как ER, изменяющего локальную организацию хроматина, требует построения карт хромосомных контактов в геноме человека, опосредованных рецептором ERα, с помощью новых технологий секвенирования – таких как ChIA-PET.

1.5. ФАКТОРЫ ПОДДЕРЖАНИЯ ПЛЮРИПОТЕНТНОСТИ В ЭМБРИОНАЛЬНЫХ СТВОЛОВЫХ КЛЕТКАХ

Задачи исследования транскрипционных факторов, связанных с поддержанием плюрипотентности эмбриональных стволовых клеток

В данном разделе рассмотрены проблемы исследования транскрипционных факторов, ответственные за поддержание клеток в плюрипотентном состоянии. Необходимость исследований молекулярных механизмов поддержания плюрипотентности связана с медицинской значимостью проблемы - эмбриональные

стволовые клетки (ЭСК) человека потенциально могут служить неисчерпаемым источником для деривации клинически ценных специализированных клеток для регенеративной медицины и клеточной терапии.

На молекулярном уровне процесс дифференцировки клеток сопровождается изменением экспрессии генов, статуса метилирования генома, модификацией белков. Выявление набора транскрипционных факторов определяющих дифференцировку клетки или сохранение плюрипотентного состояния, определение генов мишеней таких факторов и регуляторной генной сети их взаимодействий - актуальная задача фундаментальной биологии и медицины.

Анализ транскрипционных факторов плюрипотентности имеет и практическое медицинское значения, связанное с терапией стволовых клеток. Прямое репрограммирование соматических клеток с помощью набора определенных факторов впервые было сделано в 2006 году японскими учеными Такахаши и Яманака [406]. В ставшем уже классическом эксперименте было показано, что достаточно экспрессии только четырех генов, кодирующих транскрипционные факторы Oct3/4, Sox2, KLF4, и c-Myc (так называемый «коктейль Яманака») для того чтобы фибробласты кожи мыши перешли в плюрипотентное состояние. Выполнялись исследования профилей связывания девяти транскрипционных факторов в ЭСК мыши [407]. Серия последующих работ позволила уточнить минимальный набор факторов репрограммирования, найти варианты повышения его эффективности, исследовать сайты связывания этих транскрипционных факторов в масштабе генома для мыши как модельного организма [3, 39-41, 54] и для человека [19, 42].

1.5.1. Эмбриональные стволовые клетки

В процессе индивидуального развития организма млекопитающих, клетки эмбриона проходят множество стадий, постепенно теряя способность к дифференцировке: от тотипотентной зиготы через стадию плюрипотентных клеток внутриклеточной массы (ВКМ) бластоцисты, к мультипотентным стволовым клеткам и, наконец, к терминально дифференцированным клеткам [408]. При нормальном развитии процесс дифференцировки клеток сопровождается изменением экспрессии генов, статуса метилирования генома, модификацией гистоновых белков.

Первые попытки де-дифференцировки клеток in vitro были предприняты еще в середине прошлого века на амфибиях, а позже и на млекопитающих. Было показано, что при переносе ядра соматической клетки в энуклеированный ооцит с низкой

эффективностью происходит восстановление дифференцировочного потенциала соматической клетки до плюрипотентного состояния in vivo [409].

Эксперименты по слиянию ЭСК с соматическими клетками позволяют получить плюрипотентные гибридные клетки, способные принимать участие в формировании организма химерных животных [410, 411]. Эти исследования показали принципиальную возможность репрограммирования генома соматической клетки до плюрипотентного состояния in vitro, и позволили предположить, что в клетке уже содержатся все необходимые факторы для репрограммирования. Однако, набор факторов, необходимых для обратной дифференцировки клеток, оставался неизвестным до экспериментов Яманака [406, 412]. Было показано, что достаточно экспрессии четырех генов, кодирующих транскрипционные факторы Oct3/4, Sox2, KLF4, и c-Myc для того чтобы фибробласты кожи мыши перешли в плюрипотентное состояние [406, 408]. Полученные клетки были названы индуцированными плюрипотентными стволовыми клетками (ИПСК) или iPSC (induced plutipotent stem cells). По своим свойствам они оказались практически идентичны эмбриональным стволовым клеткам (ЭСК). Как и ЭСК, ИПСК продолжают симметрично делиться в присутствии факторов, обеспечивающих самоподдержание, и при отсутствии сигналов дифференцировки. В то же время, сменив условия культивирования, можно получить контролируемую дифференцировку ЭСК и ИПСК в клетки трех зародышевых листков - эктодермы, энтодермы и мезодермы.

Получив соматические клетки пациента (например, клетки кожи), можно из них сделать ИПСК, получить пораженный тип клеток, и попробовать подобрать индивидуальные лекарственные средства для устранения причин патологии. Более того, можно устранить генетическую причину заболевания и использовать клетки для трансплантации. О важности этих исследований говорит присуждение С. Яманака и Дж. Гердону Нобелевской премии по медицине и физиологии за 2012 год.

Однако, до сих пор остается нерешенным целый ряд вопросов фундаментального характера: как именно происходит репрограммирование, какие процессы обеспечивают возвращение клетки в плюрипотентное состояние, какие транскрипционные факторы и гены-мишени их воздействия важны для поддержания плюрипотентного состояния клетки.

1.5.2. Транскрипционные факторы плюрипотентности и репрограммирование

С момента открытия С. Яманака в 2006 году индуцированной плюрипотентности [406] опубликовано более пяти тысяч статей, описывающих альтернативные

комбинации транскрипционных факторов, новые способы получения ИПСК у разных организмов, включая человека, предлагающих различные способы доставки (вирусный, плазмидный, транспозонный, белковый, РНКовый), использующих различные исходные соматические клеточные типы. Репрограммированные клетки необходимы как при создании моделей in vitro широкого спектра заболеваний человека, так и при скрининге лекарственных препаратов, что требует продолжения исследований.

В 1957 году Конрад Уоддингтон предложил схематичную аналогию процесса развития животного организма из оплодотворенной яйцеклетки [413, 414] (см. рисунок 1.12). Эта аналогия известна в науке как «морфогенетический ландшафт Уоддингтона».

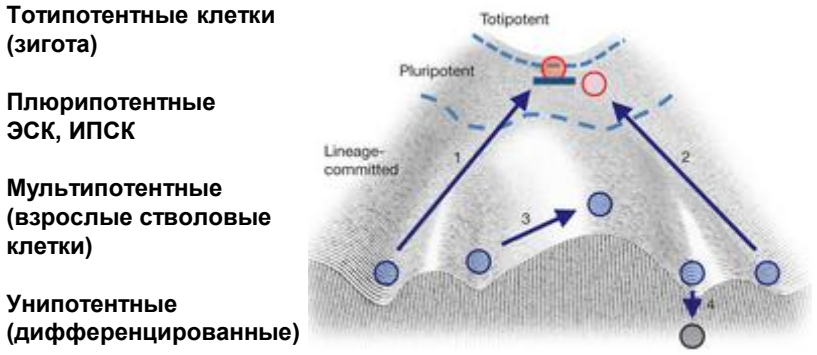


Рис. 1.12. Репрограммирование и морфогенетический ландшафт Уоддингтона [414].

В процессе развития каждая клетка проходит путь от изначального недифференцированного состояния, характерного для оплодотворенной яйцеклетки (зиготы) и клеток раннего эмбриона, до специализированных клеточных форм. Наиболее «многофункциональную» клетку — зиготу — называют тотипотентной (от лат. totalis – всеобщий), то есть, дающей начало всем другим типам клеток. По мере развития, клетки становятся все более специализированными, у дочерних клеток потенциал снижается (условный «шарик» скатывается в одну из «долин» ландшафта), тотипотентность теряется [414]. Из дочерних клеток возникают многие типы клеток. Это состояние принято называть плюрипотентностью (от лат. pluris – многое). На рисунке плюрипотентное состояние обозначено красным цветом. Соматические клетки (клетки тела) уже растратили свои потенции, они находятся в «долинах» ландшафта (обозначены синим цветом на рисунке 1.12). Некоторые соматические клетки можно перепрограммировать до плюрипотентного состояния. По мере делений они снова «скатятся» вдоль путей развития.

Роль транскрипционных факторов поддержания плюрипотентности, модель регуляторного контура эмбриональных стволовых клеток, основанная на действии OCT4, SOX2, NANOG [415, 416] представлена на рисунке 1.13.

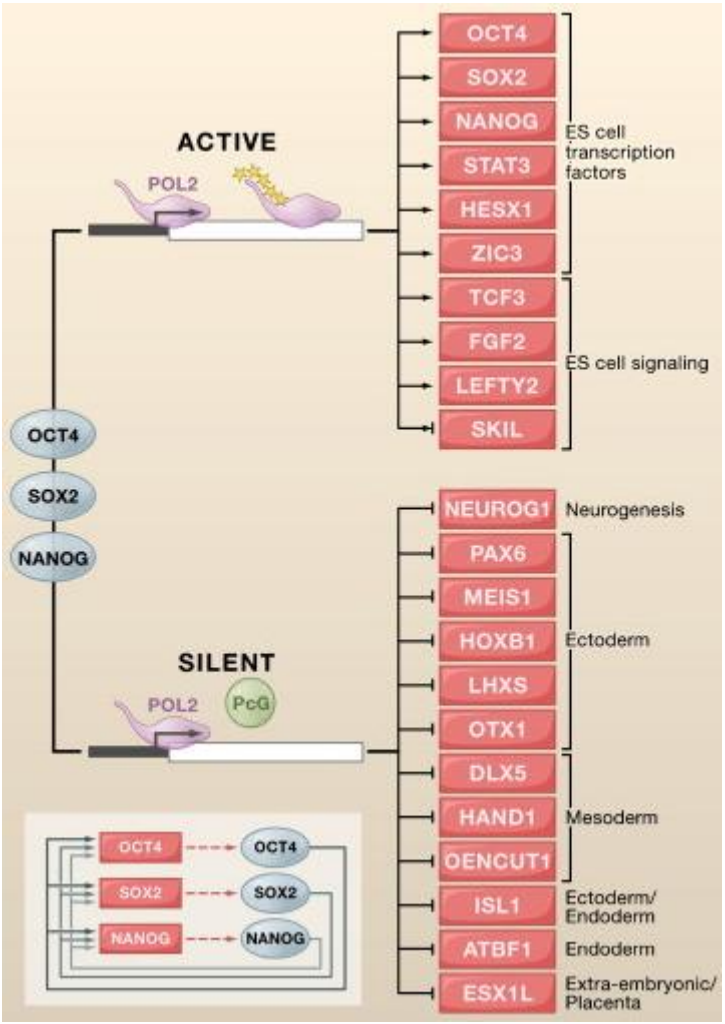


Рис. 1.13. Модель регуляторного контура эмбриональных стволовых клеток, основанная на действии факторов OCT4, SOX2 и NANOG [416].

Рисунок 1.13 показывает группы генов-мишеней OCT4, SOX2 и NANOG, активируемых в ЭСК (сами OCT4, SOX2 и NANOG, а также STAT3, ZIC3, HESX1 и другие), и репрессируемых этими факторами генов дифференцировки (NEUROG1, PAX6, MEIS1, HOXB1 и др.).

ЭСК мыши, используемые в качестве модели, были впервые изолированы в 1981 году из бластоцистов мыши [417]. Возможность получения трансгенных ЭСК мыши через гомологичную рекомбинацию привела к созданию генетически модифицированных животных - важнейшей вехе в биологических исследованиях [418]. Поддержание самообновляющегося состояния ЭСК мыши требует цитокина LIF (фактора, ингибирующего лейкемию - Leukemia Inhibitory Factor). Связывание LIF с его рецептором активирует STAT3 через фосфорилирование [419]. Одного только LIF не достаточно для поддержания ЭСК, поскольку поддержание клеток требует присутствия сыворотки новорожденных телят (fetal calf serum). Белки BMP (морфогенетические белки кости) оказались ключевым фактором сыворотки, действующим в паре с LIF для

усиления самообновления и плюрипотентности ЭСК [420]. Связывание BMP4 с его рецепторами переключает фосфорилирование Smad1 и активирует экспрессию членов семейства генов ингибиторов дифференциации *Id* (inhibitor of differentiation). Поскольку ЭСК, сверхэкспрессирующие гены *Id*s, могут самообновляться в отсутствии BMP4, было определено, что индукция экспрессии *Id* вносит критически важный в путь передачи сигнала BMP/Smad. Следовательно, сигнальные пути LIF и BMP играют центральную роль в поддержании свойств и фенотипа плюрипотентной стволовой клетки.

В целом большое значение имеет задача повышения эффективности репрограммирования, которая может быть повышено путем добавления химических компонент, таких как ингибиторы ДНК-метилтрансферазы, гистон деацителазы, киназы митоген-активированного протеина (MAPK) и киназы-3 гликоген синтазы (GSK3) [421-423]. Хотя ИПСК имеют такую же морфологию и экспрессируют молекулярные маркеры, похожие на ЭСК, их способности и степень вклада в химеризм сильно варьируют [422, 424]. ИПСК не полностью воспроизводят свойства эмбриональных стволовых клеток [425]: отмечается различие в качестве различных линий ИПСК.

Кроме сигнальных путей, которые чувствительны к присутствию внешних ростовых факторов в окружении, транскрипционные факторы также существенны для задания недифференцированного состояния ЭСК. Транскрипционный фактор Oct4, кодируемый геном *Pou5f1*, это POU домен-содержащий ТФ, существенный для ЭСК и раннего эмбрионального развития [417, 426, 427]. Oct4 взаимодействует с Sox2 (HMG-содержащий ТФ), полногеномное картирование сайтов связывания OCT4 и SOX2 в ЭСК человека показало, что они совместно воздействуют на множество генов [428]. Цис-регуляторный элемент, с которым связывается комплекс Sox2-Oct4 состоит из соседних элементов *sox* (5′-CATTGTA-3′) и *oct* (5′-ATGCAAAT-3′), представленных соответствующими консенсусами [429].

В серии работ, включая известную публикацию Яманака [406], показано, что Oct4 и Sox2, вместе с c-Myc и Klf4, достаточны для репрограммирования фибробластов в индуцированные плюрипотентные стволовые клетки (ИПСК, или iPSC), которые функционально похожи на ЭСК [430-432]. Следовательно, эти ТФ могут выполнять доминирующую роль в реконструировании транскрипционной регуляторной сети ЭСК. Nanog - это третий хорошо изученный ТФ в ЭСК. Nanog - это гомеодомен-содержащий ТФ, который может поддерживать состояние плюрипотентности в ЭСК даже в отсутствие LIF [433, 434]. Показано, что KLF (Krüppel-like factor) белки имеют GC-богатый сайт связывания подобный сайту связывания белков семейства Sp1 [435].

Klf4 также как и близкие члены того же семейства ТФ Klf2 и Klf5 важны для самообновления ЭСК мыши [436]. Другая группа [437], также показала, что Klf5 важен для поддержания ЭСК, подтверждая тем самым результаты [436]. Недавнее исследование [438] показало, что клеточные пути Oct4 и LIF-Stat3 активируют Klf2 и Klf5, соответственно, для поддержания самообновления ЭСК. Хотя все три Klf белка вовлечены в самообновление ЭСК, фактически есть избыточность в этих трех белках Klf, поскольку только тройной нокаут Klf2, Klf4 и Klf5 в ЭСК мыши индуцирует явный дифференцированный фенотип [436]). В соответствии с этим результатом, Nakagawa и соавторы (2008) показали, что Klf2 и Klf5 могут заменять Klf4 в репрограммировании соматических клеток [439]. Интересно отметить, что кроме способности генов из тех же семейств, что и Klf4, Sox2, c-Myc заменять их аналогов в репрограммировании [439], несколько факторов Яманака могут быть замещены другими неродственными транскрипционными факторами [40, 424].

Для поддержания ЭСК требуются и другие транскрипционные регуляторы. Начата серия работ по идентификации новых компонент транскрипционной регуляторной сети, необходимых для поддержания плюрипотентности. С помощью генетических исследований было показано, что ТФ Esrrb и Zfx регулируют самообновление ЭСК [429, 440, 441]. Известно, что Smad1 и STAT3 ключевые компоненты сигнальных путей, опосредованных BMP и LIF. Для поддержания плюрипотентности важны и другие ТФ. Ранее было показано, что ТФ Esrrb находится в том же белковом комплексе, что и Nanog [442]. Интересно отметить, что сайты связывания орфанного (т.е. не имеющего лигандов) рецептора Esrrb (ранее называемый ERR2, от estrogen-receptor-related) имеет сайты связывания, близкие по строению к сайтам связывания рецептора эстрогенов, но не активируется при обработке клеток эстрадиолом [443]. ТФ Tcfcp2l1 (transcription factor CP2-like 1), принадлежащий к семейству транскрипционных факторов CP2 имеет повышенную регулируемую экспрессию в ЭСК [440], но сайты связывания не были охарактеризованными. Фактор E2F1 известен из-за его роли в регуляции прогрессии клеточного цикла, показана ассоциация участков его связывания с промоторными районами генов [444]. Показано участие ТФ Klf4 и Myc в поддержании недифференцированного состояния ЭСК [436, 445].

1.5.3. Эффективность репрограммирования и дополнительные факторы

Технология получения ИПСК интенсивно развивается в последние годы [446]. Ранее использовались векторы на основе ретровирусов, которые могли

неконтролируемо встраиваться в геном, сейчас используют аденовирусы или другие вирусные векторы, не встраивающиеся в хромосомы, а также РНК, белковые транскрипционные факторы и эписомальные плазмиды. Удалось снизить число репрограммирующих факторов с четырех до одного - так, нейрональные стволовые клетки мыши превращаются в ИПСК введением одного только фактора Oct4 [447]. Идет активный поиск способов трансдифференцировки - преобразования одного типа клеток в другой, минуя стадию стволовых клеток. Еще в 1987 году было выполнено преобразование клеток фибробласты в миобласты активацией гена MyoD [448]. Показано, что ядерный рецептор Nr5a2, связывающийся с проксимальным промотором и проксимальным энхансером гена *Pou5f1* [449], также вовлечен в поддержание ЭСК мыши [450, 451].

Преимущество ИПСК перед ЭСК состоит в том, что они могут быть получены из клеток взрослого организма, а не из эмбриона. В практическом применении, инъекция пациенту его же собственных ИПСК, обычно приводит к иммунной реакции [452]. Трансплантация плюрипотентных стволовых клеток может продуцировать образования, известные как тератомы - гетерогенные скопления зародышево-подобных тканей. Вкрапление в пересаживаемые дифференцированные клетки даже нескольких недифференцированных клеток потенциально достаточно для формирования тератом [453]. Данные о том, что аутологичные (собственные) полученные из ИПСК тератомы иммуногенны для пациента, подняли новый класс проблем исследования терапевтического потенциала [453, 454]. Недавние исследования иммуногенности ИПСК различных тканей мыши показали минимальную иммуногенность in vitro дифференцированных ИПСК [455].

Полагают, что, некоторые клетки, дифференцированные из ИПСК и ЭСК, пересаженные пациенту, продолжают синтезировать эмбриональные изоформы белков и неадекватно интерпретируют сигналы окружающих их клеток [452, 456]. Образование тератомы из ИПСК может быть вызвано низкой активностью фермента Pten, который способствующей сохранению небольшой части популяции онкогенных клеток карциномы, инициирующих тератомы [457].

Показано, что пересадка ядер соматических клеток, слияние клеток, экспрессия линие-специфичных факторов способны индуцировать изменения в развитии клеток различных соматических типов, таким образом, есть возможность прямого перепрограммирования клеток различного происхождения [458]. Найдена комбинация минимального количества генов (Ngn3, Pdx1 и Mafa), с помощью которых можно перепрограммировать дифференцированные клетки взрослого организма в клетки,

проявляющие свойства эндокринных клеток поджелудочной железы [459]. Были найдены способы преобразования экзокринных клеток поджелудочной железы в эндокринные [459], клеток фибробласт - в кардиомиоциты [460]. Показан пример превращения друг в друга клеток разных зародышевых листков - мезодермальных фибробластов в эктодермальные нейроны [458, 461]. Совместная экспрессия трех транскрипционных факторов, Brn2 (также известный как Pou3f2), Ascl1 и Myt1l, может эффективно преобразовывать фибробласты мыши в функциональные индуцированный нейронные клетки. Таким образом, соматические клетки человека не из нейронов, также как и плюрипотентные стволовые клетки, могут быть преобразованы непосредственно в нейроны определенными линие-специфичными транскрипционными факторами [461].

Способы доставки репрограммирующих факторов в ядро можно подразделить на вирусные и невирусные, а также на связанные с интеграцией в геном векторов, содержащих факторы репрограммирования, и действующие без интеграции. Более распространены вирусные векторные системы. Для доставки генов в клетку обычно используют ретровирусы, содержащие одноцепочечную РНК, в частности лентивирусы [462]. С помощью обратной транскрипции на РНК вируса синтезируется линейная двухцепочечная ДНК, которая затем интегрируется в двухцепочечную ДНК генома клетки-хозяина. По сравнению с вирусными векторами, не-вирусные векторы потенциально менее иммуногенны. Невирусным способом является прямая доставка в клетку синтетической мРНК, кодирующей четыре канонических фактора Яманака: KLF4, c-MYC, OCT4, и SOX2 [463].

Привлекательным методом невирусной доставки генов является использование транспозонов. Разработано несколько транспозонных систем для транспортировки генов в клетки - Sleeping Beauty («Спящая красавица») и PiggyBack [464]. Разработаны эффективные методики получения ИПСК мыши и человека введением в соматические клетки векторов на основе PiggyBack [465]. Используются также векторные системы на основе эписомных плазмид [466]. Возможно репрограммирование с помощью низкомолекулярных соединений [467, 468]. Отмечается связь репрограммирования с активностью миРНК [469].

Известно, что по ряду характеристик, таких как глобальный паттерн экспрессии генов [470], метилирования ДНК и распределения модификаций гистонов [471], плюрипотентные клетки, ЭСК и индуцированные плюрипотентные клетки, существенно отличаются от дифференцированных клеток. Различия по этим параметрам, указывают, что плюрипотентные клетки могут иметь особую,

отличающуюся от дифференцированных клеток, пространственную организацию генома, что подтверждается недавними исследованиями [472]. Подходы к анализу пространственной организации хромосом с помощью методов высокопроизводительного секвенирования представлены в следующем разделе обзора литературы.

1.5.4. Задачи по определению сайтов связывания факторов в ЭСК

Для исследования молекулярных механизмов поддержания плюрипотентности с помощью разрабатываемых в данной работе программ требуется компьютерный анализ данных экспериментов ChIP-seq в эмбриональных стволовых клетках (ЭСК) мыши для транскрипционных факторов (Oct4, Nanog, Sox2, Klf4, Tbx3, Eset, Nr5a2, Smad2).

Необходима реконструкция полногеномного распределение сайтов связывания этой группы транскрипционных факторов в геноме мыши, исследование совместной геномной локализация сайтов связывания транскрипционных факторов Oct4, Nanog, Sox2 и Klf4, относящихся к ключевым регуляторам плюрипотентности.

Для уточнения моделей действия транскрипционного фактора в условиях активации и подавления экспрессии под действием внешних факторов интерес представляет компьютерный анализ данных экспериментов ChIP-seq для транскрипционного фактора Smad2 в культуре ЭСК мыши, построение геномных карт связывания ССТФ Smad2. Для изучения различий в наборах генов мишеней этот ТФ в условиях активации и ингибирования необходим анализ микрочиповых данных и выявление нуклеотидных мотивов транскрипционных факторов, связывающихся в геномных окрестностях сайтов Smad2.

Для исследования функционирования транскрипционных факторов плюрипотентности в геноме человека интерес представляет реконструкция списка генов, ответственных за поддержание плюрипотентности в ЭСК человека в эксперименте с последовательным нокаутом транскриптов, и построение генной сети их взаимодействия. Также необходим компьютерный анализ данных экспериментов ChIP-seq для транскрипционных факторов OCT4, NANOG, SOX2 и PRDM14 в ЭСК человека, определение нуклеотидных мотивов связывания и кластеризация групп сайтов связывания транскрипционных факторов плюрипотентности в геноме человека.

1.6. ПРОСТРАНСТВЕННЫЕ КОНТАКТЫ ХРОМОСОМ В ЯДРЕ

1.6.1. Проблема исследования контактирующих участков хромосом

Структура хроматина и архитектура интерфазного ядра являются важнейшими элементами регуляции основных генетических процессов: транскрипции и репликации. Проблема исследования регуляторных районов транскрипции уже не в линейном, а в трехмерном, пространственном представлении исключительно интересна и быстро развивается только в последние годы [472]. Помимо традиционных молекулярно-биологических методов, таких как микроскопия, гибридизация in situ, появились новые подходы, связанные с полногеномным секвенированием и иммунопреципитацией хроматина [473]. В данном разделе приведен обзор проблем исследования состояния хроматина в интерфазном ядре, представлены методы Hi-C и ChIA-PET для определения хромосомных контактов с помощью секвенирования.

Компактизация хроматина и регуляция транскрипции

Интенсивное развитие флюоресцентной микроскопии и флюоресцентной *in situ* гибридизации (FISH) в последние десятилетия сделало возможным визуализировать трехмерную организацию хромосом в интерфазном ядре уровне – анализе глобальной 3-х мерной архитектуры генома [64, 472]. Было показано, что хромосомы на любой стадии клеточного цикла представляют собой более или менее компактные структуры, занимающие отдельные, неперекрывающиеся друг с другом, области ядерного пространства, получившие название «хромосомные территории» [474].

Известно, что хромосомы в интерфазном ядре занимают определенные, не перекрывающиеся друг с другом области, формируя так называемые хромосомные территории, причем более плотно упакованный гетерохроматин располагается преимущественно в периферийной и приядрышковой зонах ядра [475]. Репликация деконденсированного эухроматина и плотно упакованного гетерохроматина разнесена по времени. Репликация деконденсированного эухроматина происходит в начале S-фазы, в то время как зоны конденсированного хроматина реплицируются в ее конце.

Хромосомные территории распределены в ядре неслучайным образом, то есть территории отдельных хромосом предпочтительно локализуются в определенном районе ядра: либо у периферии, либо в центре. Зачастую обогащенные генами хромосомы располагаются ближе к центру ядра [476, 477]. Было показано, что в некоторых клеточных типах организация хромосомных территорий в интерфазном ядре претерпевает значительные изменения. Например, у млекопитающих, ведущих ночной

образ жизни, в фоторецепторных клетках неактивные гетерохроматизированные районы хромосом перемещаются в центральную область ядра [64, 478].

Рисунок 1.14 представляет схему расчета хромосомных контактов в ядре для хромосом генома – двумерная карта (матрица контактов), полученную по координатам контактирующих фрагментов и визуализацию плотности хромосомных контактов, рассчитанных по таким матрицам, в форме тепловой карты (термокарты). Матрица контактов симметрична, цветом выделяются районы более частых контактов, полученные в эксперименте.

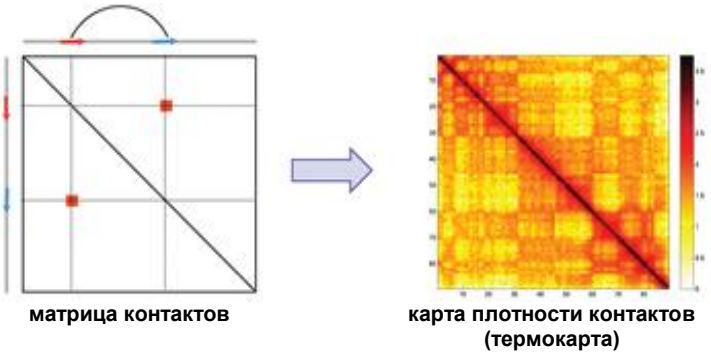


Рис. 1.14. Схема компьютерного представления хромосомных контактов в ядре клетки для хромосом генома – двумерная карта (матрица контактов) (слева) и тепловая карта плотности контактов (справа) [479].

Методы определения хромосомных контактов с помощью микроскопии и FISH

Исследования с использованием FISH проб к специфическим хромосомным локусам представляют убедительные доказательства организации ядерного материала по модели хромосомных территорий. Однако исследование пространственной структуры ядра методом FISH имеет ряд серьезных ограничений: 1) в эксперименте можно визуализировать лишь небольшое количество конкретных хромосомных локусов; 2) исследование можно провести на небольшом (порядка нескольких сотен) количестве клеток; 3) пространственное разрешение метода ограниченно. Так, для того чтобы два локуса в ядре клетки были найдены, в геноме они должны быть разделены не менее чем 100 тысячами пар нуклеотидов (т.п.н.) [480, 481], и даже применение высокоразрешающей микроскопии имеет свои ограничения, поскольку необходимость стадии денатурации ДНК ставит вопрос о сохранности наноразмерных хроматиновых структур при приготовлении препаратов для FISH. Поэтому все большое значение приобретают исследования тонкой организации структуры хроматина с помощью молекулярно-биологических подходов.

1.6.2. Методы определения хромосомных контактов с помощью секвенирования: 3C и Hi-C

Появившийся недавно метод Hi-C позволяет оценивать полноту репрограммирования на принципиально новом уровне – анализе глобальной трехмерной архитектуры генома [64, 482]. Трехмерная организация хромосом в ядре имеет большое значение для многих клеточных процессов, включая регуляцию экспрессии генов, репликации ДНК, и структуры хроматина [483]. Хотя упаковка хроматина через образование нуклеосомных нитей относительно хорошо изучена, мало что известно об укладке на наднуклеосомных уровнях организации хроматина.

Первым методом, позволяющим исследовать пространственное взаимодействие двух специфических районов генома без привлечения микроскопии, был метод 3C (Chromosome Conformation Capture – захват конформации хромосом) [484]. В отличие от методик микроскопии, 3C и другие методы, разработанные на его основе, не фиксируют конкретное событие взаимодействия в пределах одной клетки, а измеряют вероятность взаимодействия двух районов генома в большой (10^5–10^6) популяции клеток. Метод 3C основан на принципе лигирования сближенных в пространстве молекул ДНК и включает 4 этапа. Первый этап – фиксация клеток формальдегидом для сохранения нативной трехмерной структуры ядра (рис. 1.15, а).

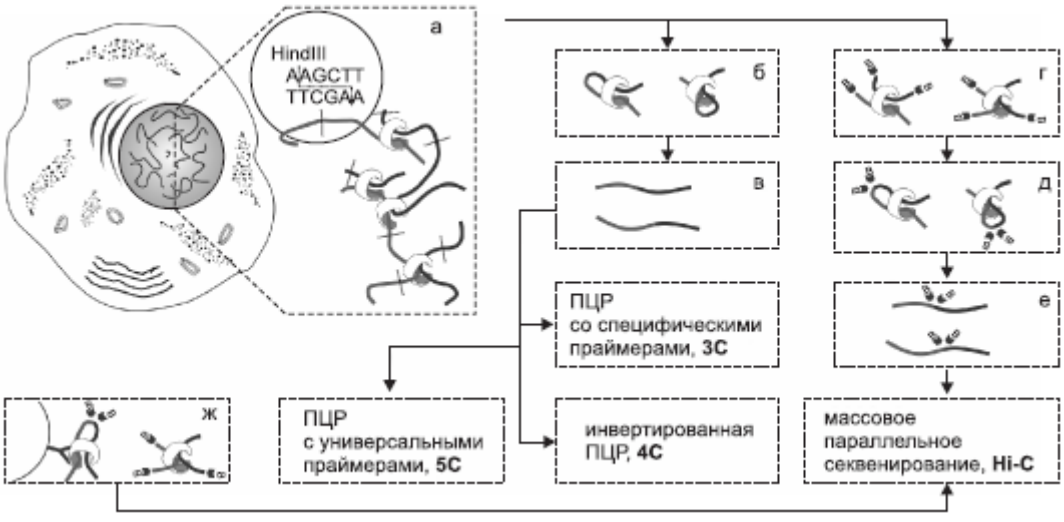


Рис. 1.15. Методы, основанные на технологии 3C.
а – стадия фиксации клеток формальдегидом и обработки рестриктазой;
б – лигирование сближенных в пространстве молекул ДНК;
в – выделение ДНК и последующий анализ библиотеки методами ПЦР;
г – заполнение липких концов ДНК биотинилированными нуклеотидами;
д – лигирование сближенных в пространстве молекул ДНК;
е – обогащение библиотеки продуктами лигирования и массовое параллельное секвенирование библиотеки;
ж – иммунопреципитация хроматиновых комплексов и массовое параллельное секвенирование, метод ChIA-PET [64].

Формальдегид фиксирует белок-белковые, белок-ДНК и белок-РНК-взаимодействия за счет образования ковалентных связей между первичной аминогруппой белка и нуклеиновой кислотой. В 3C экспериментах чаще всего применяется именно формальдегид, поскольку поперечные сшивки, которые он образует, имеют наименьший размер (2 ангстрема) среди других фиксирующих агентов. Эта особенность позволяет повысить пространственное разрешение метода, кроме того, фиксация формальдегидом обратима [484-487]. После того как взаимодействующие молекулы ДНК оказываются сшитыми с белками, ДНК фрагментируется с помощью рестриктаз (рис. 1.15, а).

После этого проводят лигирование в условиях сильного разбавления (рис. 1.15, б). В таких условиях лигируются только концы молекул ДНК, сближенных в пространстве. Далее химерные молекулы ДНК выделяются и очищаются; создается библиотека попарно взаимодействующих молекул ДНК. Относительное обогащение в библиотеке специфических районов генома, лигированных друг с другом, отражает вероятность взаимодействия этих районов в трехмерном пространстве ядра в популяции клеток. Типичная 3C библиотека содержит огромное (до 10^{11}) количество уникальных пар взаимодействующих районов [24].

При необходимости оценить взаимодействие двух конкретных районов генома, достаточно воспользоваться методом количественной ПЦР, подобрав один из праймеров, гомологичный одному району, а второй – другому [488]. Получение продуктов амплификации такой пары праймеров при анализе 3C-библиотеки будет свидетельствовать о близком пространственном расположении выбранных районов генома в ядре клетки (рис. 1.15, в).

Для поиска последовательностей ДНК, контактирующих с выбранным участком генома, можно воспользоваться методами 4C и 5C. 4C (Circularized Chromosome Conformation Capture – замкнутый захват конформации хромосом) представляет собой объединение метода 3C и метода инвертированной ПЦР [21, 489]. 5C (Carbon-Copy Chromosome Conformation Capture – захват конформации хромосом в копиях) – объединение методов 3C и мультиплексной ПЦР с предварительным лигированием адаптерных последовательностей [490]. 5C позволяет проводить одновременный поиск районов, контактирующих с несколькими выбранными участками генома.

Также стоит упомянуть о методах ChIP-loop [491] и ChIA-PET [492], которые позволяют выявить участки генома, контактирующие с помощью специфических белков, например, транскрипционных факторов или белков инсуляторов. Эти методы основаны на принципах иммунопреципитации и лигирования сближенных в

пространстве молекул ДНК [21] (рис., ж). Подробное описание 3С методов приведено в обзоре [493].

Особенно перспективным на сегодняшний день является метод Hi-C (High-throughput Chromosome capture), который позволяет определять пространственную структуру хроматина в масштабе всего генома с очень высоким разрешением [482, 484] и позволяет реконструировать карту пространственных взаимодействий ДНК в ядре клетки. Hi-C представляет собой объединение метода хромосомного захвата (Chromosome Capture), который ранее использовался для идентификации потенциальных цис-регуляторных элементов, с технологиями массового параллельного секвенирования, позволивших вести поиск взаимодействующих сайтов в масштабе всего генома. Последние работы представляют применения метода Hi-C уже для отдельных клеток [494].

Метод Hi-C представляет собой объединение технологии 3С и технологий массового параллельного секвенирования. Глубокое секвенирование теоретически позволяет установить все хроматиновые контакты, существующие в геноме. На практике количество установленных контактов сильно зависит от глубины секвенирования библиотеки [495]. Основной технической особенностью создания Hi-C библиотеки является этап обогащения библиотеки продуктами межмолекулярного лигирования. Такое обогащение достигается за счет заполнения липких концов, возникших при обработке рестриктазой фиксированного хроматина, биотинилированными нуклеотидами (рис. 1.15, г). На следующем этапе проводится лигирование по тупым концам (рис. 1.15, д). Таким образом, продукты межмолекулярного лигирования, собственно молекулы, несущие информацию о контактах ДНК, оказываются меченными биотинилированными нуклеотидами. Использование на следующем этапе магнитных частиц, покрытых стрептавидином, позволяет сконцентрировать продукты лигирования (рис. 1.15, е) и использовать их для массового параллельного секвенирования [482].

С использованием улучшенной методики Hi-C (закреплении хроматина перед стадией лигирования на поверхности магнитных частиц) удалось показать, что до половины всех межхромосомных контактов, выявленных с использованием традиционного протокола Hi-C, являются артефактами [22]. Улучшенная методика Hi-C, названная TCC (Tethered Conformation Capture – связанный конформационный захват), позволила детально исследовать характер межхромосомных взаимодействий [22]. Принципиальным открытием стало то, что межхромосомные контакты очень динамичны, т.е. в различных клетках одной и той же клеточной популяции

распределение контактов может значительно отличаться. Авторы показывают, что из всего пула межхромосомных контактов лишь 20% являются общими для любых двух клеток, взятых из популяции [22].

Было показано, что районы активно транскрибируемых генов чаще участвуют в межхромосомных контактах. По-видимому, этот факт объясняется выделением петель ДНК из хромосомной территории в область концентрации белков транскрипционного аппарата (так называемые «фабрики транскрипции»), где становится возможным контакт с другим активным районом. Предсказанные на основе данных Hi-C частоты межхромосомных контактов хорошо согласовывались с данными, полученными на основе 3D-FISH-анализа [496].

Существует ряд проблем, связанных с 3C-технологией. Поскольку формальдегид фиксирует существующую в данный момент пространственную конфигурацию ядер в большой популяции клеток, неизвестно, какая доля зафиксированных событий взаимодействия является временными флуктуациями структуры, а какая представляет собой редкие, но относительно стабильные варианты контактов [497].

Исследования паттерна репликации ДНК в ЭСК и дифференцированных клетках [498] показали, что пространственное распределение поздно- и ранореплицирующихся доменов генома с высокой точностью совпадает с трехмерной (3D) картой взаимодействующих районов ДНК [482].

Понимание того, как происходит укладка хромосом – существенный элемент в исследовании общих связей между структурой хроматина, активностью генов и функциональным состоянием клеток. Для моделирования и исследования укладки хроматина обычно используется модель бус на нитке, в которой хроматин представляет собой последовательность связанных нуклеосом в виде мономеров. При моделировании, укладка хроматиновой нити в 3D глобулу представляется за счет введения притяжения между мономерами, обусловленного растворителем, либо в результате ограничения пространства, которое может занимать нить. Наиболее известной моделью является модель «равновесной глобулы», в которой нить ведет себя случайным образом, так что далеко удаленные по цепи пары оснований ДНК могут соседствовать; при этом распределение вероятности контактов $P \sim s^{-3/2}$, где s – расстояние по цепи выраженное в мегабазах. Недавние эксперименты [482] показали, что такая вероятность обратно пропорциональна расстоянию $P \sim s^{-1}$. Была предложена модель «фрактальной глобулы» [23, 482, 499], которая приводит к этой зависимости. В отличие от равновесной глобулы, «фрактальная глобула» не имеет узлов, так что цепь

легко укладывается и разворачивается, что важно для активации и репрессии генов [482].

Метод Hi-C использует очистку продуктов лигирования с последующим массовым параллельным секвенированием. Hi-C позволяет определить взаимодействие участков хроматина во всем геноме, без выделения заранее исследуемых районов. Кратко метод состоит в следующем: клеточный материал, включая белки связанные с ДНК, фиксируется формальдегидом; ДНК энзиматически переваривается ферментом рестрикции, который оставляет свободный 5′-конец; этот 5′-участок достраивается, включая биотинилированный остаток; получившиеся тупые концы лигируются в растворе при параметрах благоприятствующих событиям лигирования между слинкованными (совместно связанными) фрагментами ДНК. Получившаяся выборка молекул ДНК содержит продукты лигирования, состоящие из фрагментов, которые исходно были в пространственной близости в ядре, маркированные биотином в точке соединения. Библиотека Hi-C содержащая такие молекулы ДНК создается путем перемешивания (shearing) ДНК и отбора фрагментов ДНК, маркированных биотином на стрептавидиновых бусах. Библиотека (полный набор) последовательностей ДНК затем анализируется с помощью массового параллельного секвенирования, определяя каталог взаимодействующих фрагментов ДНК во всем геноме.

3C-методы, и в особенности метод Hi-C, являются уникальными инструментами изучения пространственной организации ядра и архитектуры хромосом, позволяющими получить глобальную карту взаимодействующих районов генома. Однако метод Hi-C имеет и свои ограничения. Hi-C не дает информации о том, какие белки опосредуют хромосомные контакты, нет специфичности связывания, как в методах, основанных на иммунопреципитации хроматина.

1.6.3. Метод ChIA-PET

Метод ChIA-PET (Chromatin Interaction Analysis by Paired-End-Tag sequencing), использующий иммунопреципитацию хроматина позволяет определять контактирующие участки хромосом, контакты которых опосредованы белками или белковыми комплексами. Технически метод ChIA-PET основан на тех же этапах выделения контактирующих фрагментов хромосом, что и метод Hi-C, однако предназначается для решения другой проблемы, связанной с регуляцией экспрессии генов эукариот.

Исследование хромосомных контактов поднимает фундаментальный вопрос: как гены и их регуляторные районы структурно организованы для регуляции

транскрипции. Известно, что бактерии содержат структуры оперона, в которых один с одного промотора транскрибируется несколько генов в одной РНК [500]. В то же время механизм пространственной координации транскрипции в клетках эукариот остается неясным. Хотя структуры бицистронных транскриптов были описаны для геномов червей и дрозофилы [501, 502], полагают, что гены эукариот транскрибируется поодиночке, со своих собственных промоторов. Тем не менее, эксперименты микроскопии, наблюдение флюоресценции in situ за последнее десятилетие дают основания предполагать, что активная транскрипция не распределена равномерно в ядре клетки, но сконцентрирована дискретно в больших пространственных участках (локусах) в ядре клеток млекопитающих, что дает основания предположить пространственную сближенность районов транскрипции в так называемых «транскрипционных фабриках» [503], когда участки хромосом, содержащие транскрибируемые гены, комплекс РНК-полимеразы II и другие белковые компоненты транскрипционной машины физически сближены. Такой общей теории не хватает молекулярных и структурных деталей, что оставляет вопросы о координации транскрипции генов в клетках млекопитающих.

Известно, что геном млекопитающих плотно упакован в конформации высокого порядка в пространстве ядра микронного размера. Следовательно, трехмерная (3D) организация должна вносить важный вклад в механизмы регуляции транскрипции и ее глобальной координации [504]. Метод захвата хромосом, 3C (Chromosome Conformation Capture) и подобные техники [505] вместе с традиционными методами ранее были использованы для демонстрации того, что взаимодействия хроматина могут регулировать транскрипционное и эпигенетическое состояние [506]. Тем не менее, такие исследования были либо ограничены отдельными доменами генома, либо имели низкое разрешение и не описывали функциональные детали. Таким образом, необходим глобальный метод высокого разрешения для описания функциональных взаимодействий хроматина и понимания основополагающих принципов архитектуры геномной организации высокого порядка относительно регуляции транскрипции.

Был разработан метод анализа взаимодействий хроматина с помощью секвенирования парных концов ДНК (Chromatin Interaction Analysis by Paired-End-Tag sequencing - ChIA-PET) для полногеномного исследования взаимодействий хроматина, связанного со специфическими белковыми факторами [21]. С помощью иммунопреципитации исследуемого фактора вместе с ассоциированными с ним фрагментами ДНК, с последующим лигированием в растворе удаленных фрагментов

ДНК, закрепленных вместе в отдельном комплексе хроматина изучалась ассоциация регуляторных районов генов через их нелинейные, дистальные взаимодействия.

В методе ChIA-PET, так же, как для ChIP-seq, сначала выполняется иммунопреципитация хроматиновых комплексов, связанных с исследуемым белком, затем фрагменты ДНК разделяются ультразвуком (рис. 1.16).

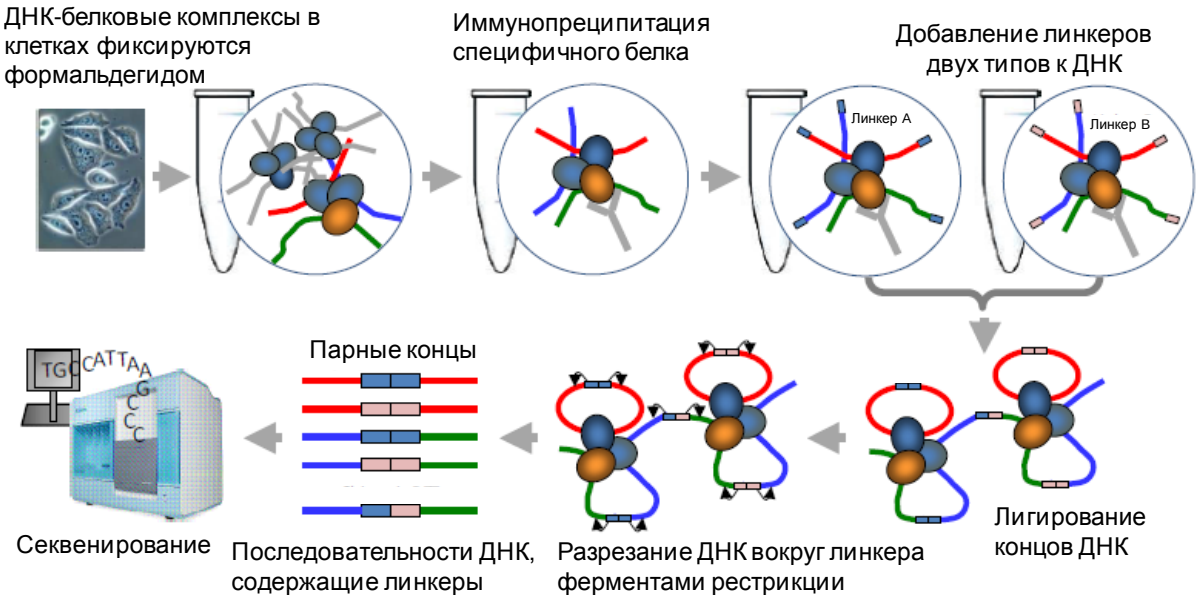


Рис. 1.16. Схема метода ChIA-PET [21]. Синим, красным и зеленым цветом обозначены последовательности ДНК, контактирующие с белками, которые затем лигируются. Парные концы вместе с линкером направляются на секвенирование.

Фрагменты ДНК из разделенных ультразвуком, прошедших иммунопреципитацию хроматиновых комплексов обрабатываются через лигирование линкеров (на свободные концы ДНК), получаются парные концы (PET). Затем выполняется разрезание ДНК вокруг связанных линкеров с помощью ферментов рестрикции. ДНК секвенируется с концов, образуя пары прочтений. Далее пары прочтений ДНК картируются на референсную последовательность генома, строится таблица парных контактов. Полученная таблица контактов (пар координат в геноме) обрабатывалась с помощью разработанной автором компьютерной программы, для выделения статистически значимых участков контактов, анализа пересечений координат с установленным ранее расположением ССТФ и районов модификаций хроматина.

По методу ChIA-PET были получены карты контактов на хромосомах для транскрипционного фактора ERα – рецептора эстрогенов [21]. Расположение хромосомных контактов в геноме человека, полученное в этой работе, было проанализировано автором диссертации с помощью разработанных компьютерных программ интеграции с полученными ранее данными по методам ChIP-PET и ChIP-seq.

Данные о хромосомных контактах, полученные с помощью секвенирования ChIA-PET, независимо были проверены для отдельных геномных локусов с помощью экспериментов по технологии 3C (Chromosome Conformation Capture) и флуоресцентной гибридизации *in situ* (FISH) - таким образом, подтверждена корректность определения контактов в полногеномном методе.

С помощью ChIA-PET определены контакты, опосредованные транскрипционным фактором CTCF [507]. Ранее были охарактеризованы дальние взаимодействия включающие α- и β-глобиновые локусы [490, 508]. Эксперименты ChIA-PET по определению хромосомных контактов в ядре клетки, опосредованных уже не отдельным транскрипционным фактором, а целым транскрипционным комплексом РНК-полимеразы II впервые были представлены в работе [12], с участием автора диссертации.

1.6.4. Постановка задач анализа данных ChIA-PET

Технология ChIA-PET определения сайтов связывания и хромосомных контактов с помощью секвенирования требует построения карт хромосомных контактов, опосредованных рецептором эстрогена ERα и комплексом РНК-полимеразы II в геноме человека. Классификация групп генов, находящихся в транскрипционных доменах, в зависимости от структуры контактов (хромосомных петель) представляет несомненный интерес. Должно быть исследовано присутствие сайтов связывания различных транскрипционных факторов, определенных с помощью технологии ChIP-seq в геноме, в участках хромосомных контактов, опосредованных комплексом РНК-полимеразы II, как пример иерархической организации регуляторных районов транскрипции. Для контактирующих районов хромосом должно быть выполнено компьютерное исследование связи с модификациями гистонов, характеризующими открытое состояние хроматина (H3K4me3, H3K9ac, H3K4me1) и другими геномными характеристиками, такими как расстояние до стартов транскрипции генов, присутствие промоторных и энхансерных районов.

ЗАКЛЮЧЕНИЕ ПО ОБЗОРУ ЛИТЕРАТУРЫ И ПОСТАНОВКА ЗАДАЧ ИССЛЕДОВАНИЯ

Представленный выше обзор подчеркивает важность разработки адекватных компьютерных моделей расположения регуляторных районов транскрипции генов в масштабе генома. Такие модели необходимы для решения широкого круга задач, связанных с картированием и разметкой регуляторных районов генов, предсказанием функциональных сайтов связывания транскрипционных факторов.

Появившиеся методы иммунопреципитации хроматина (ChIP-on-chip, ChIP-PET, ChIP-seq, ChIA-PET) с последующим массовым параллельным секвенированием позволяют исследовать сайты связывания транскрипционных факторов в масштабе всего генома. Исследование структуры хроматина в геноме на уровне отдельных нуклеосом (модификаций метилирования и ацетилирования гистонов в определенных позициях) с помощью технологий ChIP-seq определяет доступность ДНК для связывания с транскрипционными факторами и качественно дополняет описание регуляторных районов генов. Данные о роли трехмерной организации генома в регуляции экспрессии генов (удаленные энхансеры, пространственные домены), полученные с помощью технологий секвенирования ChIA-PET требуют разработки компьютерных моделей организации регуляторных районов с учетом дистальных взаимодействий.

В целом необходимо детальное теоретическое компьютерное исследование структуры регуляторных районов транскрипции генов эукариот в масштабе генома на основе анализа данных о положении сайтов связывания транскрипционных факторов, определяемых с иммунопреципитации хроматина и высокопроизводительного секвенирования (ChIP-seq).

В плане практических приложений важно компьютерное определение сайтов связывания транскрипционных факторов - регуляторов плюрипотентности NANOG, OCT4, SOX2, KLF4, PRDM14 в стволовых клетках человека и их ортологов Nanog, Oct4, Sox2, Klf4 в геноме мыши, исследование кластеров сайтов связывания различных белковых факторов в эмбриональных стволовых клетках, изучение их влияния на программы поддержания плюрипотентности. Объединение полногеномных карт сайтов связывания транскрипционных факторов с помощью программ анализа данных ChIP-seq позволит исследовать генные сети регуляции плюрипотентного состояния стволовых клеток, оптимизировать возможности репрограммирования клеток, что важно для терапии, персонализированной медицины. Большое значение для медицины

имеет задача компьютерного определения сайтов связывания транскрипционных факторов - онкогенов и супрессоров опухолей - p53, ERα, MYC, STAT1, FOXA1 в геноме человека, выявления их генов-мишеней, основанная на анализе данных ChIP-PET и ChIP-seq. Поскольку нарушение программ дифференцировки клеток, связанных с факторами самообновления и роста клеток, такими как Myc, ведет к образованию опухолей, поставленные задачи исследования транскрипционных факторов в стволовых клетках и онкогенов связаны между собой. Поиск генов и их регуляторных районов с помощью методов компьютерной биологии и анализа геномных данных может помочь в решении важных проблем фундаментальной медицины.

Следует подчеркнуть важность разработки адекватных моделей регуляции экспрессии генов на основе информации о сайтах связывания, нуклеосомной упаковке и модификациях хроматина, и данных экспрессионных микрочипов, включая распознавание сайтов связывания транскрипционного фактора и его ко-факторов в масштабе генома. Такая модель компьютерная модель распознавания сайтов должна включать данные о модификациях хроматина, полученные с помощью секвенирования.

В целом, для достижения цели исследования необходимо решение следующих практических задач:

- Разработка методов анализа данных секвенирования ChIP-seq и создание статистической модели полногеномного распределения сайтов связывания транскрипционных факторов (ССТФ).
- Компьютерная реконструкция полногеномных карт сайтов связывания транскрипционных факторов плюрипотентности c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши.
- Реконструкция распределения сайтов связывания транскрипционных факторов MYC, PRDM14, ERα, FOXA1, OCT4, NANOG в геноме человека.
- Компьютерное исследование ассоциации сайтов связывания транскрипционного фактора ERα с определенными с помощью технологии ChIP-seq маркерами хроматина, в частности, модификациями гистона H3 (H3K4me3, H3K4me1, H3K27me3, H3K9me3, H3K9ac, H3K14ac), и создание метода предсказания сайтов связывания транскрипционного фактора ERα в геноме человека на основе профилей модификаций гистонов.
- Изучение роли хромосомных контактов в регуляции транскрипции генов человека на моделях РНК-полимеразы II и транскрипционного фактора ERα на основе компьютерного анализа полногеномных данных ChIP-seq и ChIA-PET.

Методически исследование должно включать разработку и компьютерную реализацию (на языках C++, и в среде R) (1) алгоритмов анализа полногеномных профилей связывания транскрипционных факторов ChIP-seq; (2) алгоритмов анализа нуклеотидных последовательностей регуляторных районов, формируемых ССТФ; (3) алгоритма анализа полноты эксперимента ChIP-seq и ChIP-PET; (4) алгоритма определения кластеров ССТФ в геноме: (5) программ обработки данных экспрессии генов на микрочипах; (6) программ интеграции данных геномной аннотации расположения генов и профилей ChIP-seq; (7) программ анализа профилей ChIA-PET и ChIP-seq.

Практические задачи должны включать разработку Интернет-доступных программных комплексов для исследования регуляции экспрессии генов на основе данных высокопроизводительного секвенирования, разработку баз данных экспрессии генов, оценок качества микрочипов Affymetrix. Интеграция данных о хромосомных контактах в интерфазном ядре и полученных с помощью различных ChIP технологий данных о расположении регуляторных районов транскрипции генов в геноме позволит по-новому оценить проблему регуляции экспрессии, дистальной регуляции, молекулярных механизмов функционирования энхансеров. Решение задачи компьютерного анализа данных иммунопреципитации специфичных факторов и комплекса РНК-полимеразы II с учетом удаленных взаимодействий по технологии ChIA-PET должно послужить завершением представляемой работы.

Таким образом, применение компьютерных программ на наиболее полных на момент выполнения работы данных дает возможность получить новые результаты об организации регуляторных геномных последовательностей, а также задать стандарт для изучения распределения сайтов связывания транскрипционных факторов в геноме для новых данных ChIP-seq и аналогичных технологий, основанных на иммунопреципитации хроматина и секвенировании.

ПЛАН И СТРУКТУРА ИССЛЕДОВАНИЯ

В соответствии с задачами исследования, в настоящей работе кроме «Обзора литературы» представлено четыре главы: «Модели распределения сайтов связывания транскрипционных факторов в геноме», «Карты сайтов связывания по данным ChIP-seq», «Модификации хроматина и связывание транскрипционных факторов по данным ChIP-seq», «Хромосомные контакты и регуляция транскрипции в геноме человека» и Приложение.

Глава 2 «Модели распределения сайтов связывания транскрипционных факторов в геноме» содержит описание разработанных методов компьютерного анализа данных ChIP-seq, другие методические материалы и описание созданных баз данных. Глава является методической основой работы и представляет основные компьютерные алгоритмы. Основной источник информации представляют собой данные экспериментов высокопроизводительного секвенирования, сопряженного с иммунопреципитаций хроматина - ChIP-seq, ChIP-PET и ChIA-PET.

Следующая схема представляет типы данных и потоки информации, проанализированной в ходе работы.

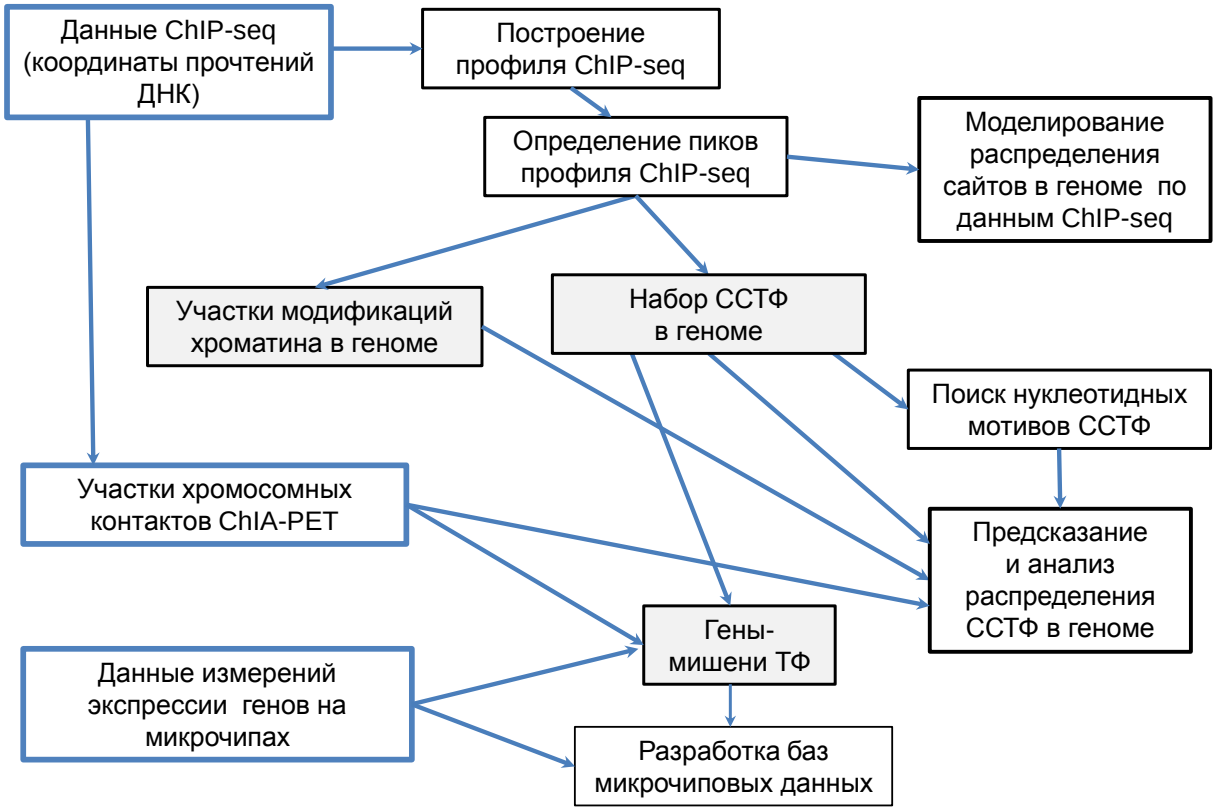


Рис. 1.17. Анализ данных в рамках диссертационной работы.

Как видно из схемы, основные типы данных - это данные ChIP-seq (координаты прочтений ДНК), данные ChIA-PET (координаты парных контактов) и данные измерения экспрессии генов на микрочипах (уровни экспрессии для нуклеотидных проб, соответствующих генам). Промежуточные данные - это координаты сайтов связывания транскрипционных факторов в геноме, координаты участков (протяженных районов) модификаций хроматина, определяемые с помощью анализа профилей связывания ChIP-seq. К промежуточным данным относятся также гены-мишени действия транскрипционных факторов, определяемые через изменений уровней экспрессии генов при активации ТФ и проксимальное расположение сайтов связывания исследуемого транскрипционного фактора. Основными результатами обработки информации являются модели распределения ССТФ в геноме по данным ChIP-seq, предсказание расположения сайтов и анализ их расположения, интеграция с другими источниками геномных данных, таких как организация хроматина, нуклеотидный контекст сайтов связывания.

Дополнительными результатами, также представленными на схеме, являются базы микрочиповых данных, включающие оценки качества микрочипов (на примере проб микрочипа Affymetrix U133 [46, 47, 49]), базы данных дифференциальной экспрессии на микрочипах, в частности для генов крысы [45, 53], база данных цис-антисенс транскриптов, связанная с оценкой качества измерения экспрессии [48, 61].

Логически исследование идет от простого к сложному - от выделения набора сайтов связывания транскрипционных факторов в геноме с помощью ChIP-seq к определению нуклеосомной упаковки и районов модификаций хроматина также с помощью высокопроизводительного секвенирования и ChIP-технологий, и к анализу уже трехмерных контактов хромосом в ядре клетки, включающих такие сайты и участки модификаций хроматина.

Компьютерное моделирование распределения сайтов связывания, выявляемых в экспериментах ChIP-seq, основано на построении полногеномного профиля ChIP-seq по координатам секвенированных фрагментов ДНК, определении пиков такого профиля. Такие модели для определения сайтов представлены в Главе 2.

Соответственно, Глава 3 «Карты сайтов связывания по данным ChIP-seq» содержит описание карт сайтов связывания транскрипционных факторов полученным по экспериментальным данным ChIP-seq. Глава 4 «Модификации хроматина и связывание транскрипционных факторов по данным ChIP-seq» содержит описание

применения разработанных компьютерных методов к исследованию модификаций хроматина и связыванию транскрипционных факторов в геноме человека [19].

Глава 5 «Хромосомные контакты и регуляция транскрипции в геноме человека» представляет исследование хромосомных контактов, полученных с помощью массового параллельного секвенирования в геноме человека по методу ChIA-PET [12, 21, 64]. Исследование сайтов связывания транскрипционного фактора - рецептора эстрогенов ERα хроматина объединяет главы 3, 4 и 5, рассматривая сначала модели сайтов связывания ERα, полученные по данным ChIP-seq, затем модификации гистонов, ассоциированные ERα, также полученные по данным ChIP-seq связанные с сайтами и далее трехмерные хромосомные контакты, опосредованные ERα.

Выполненное исследование хромосомных контактов в геноме человека по методу ChIA-PET для ERα и комплекса РНК-полимеразы II с помощью разработанных компьютерных программ [12, 21, 64] интегрирует данные ChIP-seq и ChIA-PET.

По объектам исследования - с помощью разработанных компьютерных программ обработки данных ChIP-PET и ChIP-seq были проанализированы исходные данные секвенирования и определены сайты связывания белков c-Myc, STAT1, FOXA1, ERα, PRDM14 [9, 13, 42] в геноме человека, транскрипционных факторов и регуляторов Nanog, Oct4, Sox2, Klf4, E2f1, Esrrb, CTCF, n-Myc, c-Myc, Smad1, STAT3, Tcfcp21l, Zfx, Suz12 [3]. Показано применение для генома дрожжей [51, 62, 63]. Проанализированы выборки, содержащие модификации гистонов и сайты связывания ERα, FOXA1 [13, 21, 37].

В Приложении приведены ссылки на использованные компьютерных программы, коды собственных разработанных программ и схемы алгоритмов, таблицы, содержащие координаты найденных ССТФ в геноме мыши и их кластеры, описание библиотек ChIP-seq, использованных компьютерных ресурсов и разработанных баз данных.

Глава 2. МОДЕЛИ РАСПРЕДЕЛЕНИЯ САЙТОВ СВЯЗЫВАНИЯ В ГЕНОМЕ

2.1 Компьютерные модели и базы данных. Структура Главы

Данная Глава содержит описание компьютерных моделей распределения сайтов связывания транскрипционных факторов в эукариотическом геноме. В отдельных разделах Главы представлены алгоритмы анализа профилей связывания транскрипционных факторов по данным ChIP-seq, выделения пиков профиля, оценки полноты (сатурации) эксперимента ChIP-seq. Показано применение компьютерного метода оценки полноты экспериментальных данных ChIP-seq для определения сайтов связывания транскрипционного фактора Nanog в геноме мыши.

В связи с задачами исследования представлены метода анализа данных экспрессии генов на микрочипах, оценки качества экспрессионных данных, на примере платформы микрочипов Affymetrix GeneChip. Описаны подходы к интеграции данных экспрессии генов и данных расположении сайтов связывания в регуляторных районах этих генов, а также базы данных, разработанные автором [3, 9, 16, 37, 38].

Данная Глава представляет следующие основные задачи диссертационной работы:

1. Разработка методов анализа данных секвенирования ChIP-seq и создание статистической модели полногеномного распределения сайтов связывания транскрипционных факторов (ССТФ).

2. Компьютерная реконструкция полногеномных карт сайтов связывания транскрипционных факторов c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2 и Smad2 в геноме мыши. Реконструкция распределения сайтов связывания транскрипционных факторов MYC, PRDM14, ERα, FOXA1, OCT4, NANOG в геноме человека.

После краткой постановки компьютерных задач, возникающих из технологий определения сайтов связывания транскрипционных факторов в геномах эукариот, связанных с иммунопреципитацией хроматина - ChIP-PET [9] и ChIP-seq [3], описаны используемые данные и выборки геномных последовательностей. Подробно представлены компьютерные алгоритмы анализа профилей связывания ChIP-seq в масштабе генома, анализа распределения сайтов связывания по силе связывания (по высоте пика), оценки ошибок первого и второго рода при предсказании сайтов связывания из пиков (кластеров фрагментов ДНК) профиля ChIP-seq [3, 16].

Исследование механизмов регуляции экспрессии генов должно быть дополнено собственно экспрессионными данными, полученными на тех же экспериментальных моделях - культурах клеток, что и данные иммунопреципитации хроматина. В соответствующем разделе данной Главы подробно показан контроль качества сигнала на микрочипах платформы Affymetrix, показаны приложения по экспрессии генов при раке, экспрессии не кодирующих белок транскриптов [49]. Представлен алгоритм определения генов-мишеней транскрипционных факторов по положению сайтов связывания ChIP-seq и уточнению списка генов-мишеней по экспрессии генов на микрочипах. В связи анализом механизмов регуляции экспрессии генов представлены модели регуляторных районов транскрипции, включающие цис-антисенс транскрипты, т.е. транскрипты, располагающиеся в противоположной ориентации в одном и там же геномном локусе. Описано построение компьютерной базы данных таких антисенс транскриптов в генома человека как основы исследования регуляции транскрипции генов [48].

В последнем разделе Главы описаны средства компьютерной интеграции геномных данных, разработанные в ИЦиГ СО РАН (комплекс ICGenomics) [44], показаны примеры применения для данных по экспрессии генов (База данных RatDNA) [53].

В целом, в разделах данной Главы представлены задачи анализа сайтов связывания в геноме включающие анализ специфических пиков распределения ChIP-seq, оценку общего числа сайтов в геноме, соотношение сайтов связывания с экспрессией генов на микрочипах, оценки нуклеотидных мотивов ДНК, найденных в массовых полногеномных экспериментах, описанные в работах автора [3, 9, 13, 16].

Базы данных и выборки геномных последовательностей для анализа

В работе использовались следующие данные: (1) сырые данные секвенирования участков ДНК, полученные методом ChIP-seq для факторов c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши, транскрипционных факторов MYC, FOXA1, ERα, PRDM14 в геноме человека; (2) данные экспрессии генов на микрочипах платформы Affymetrix U133 в опухолевых клетках человека, платформы Nimblegen (www.nimblegen.com/*)* в ЭСК мыши, и Illumina в ЭСК человека; (3) данные геномной аннотации RefSeq, UCSC genes, GenBank, mRNA, EST (http://genome.ucsc.edu); (4) данные секвенирования нуклеосомной ДНК для генома дрожжей *S.cerevisiae*. Использовались данные секвенирования, полученные в Геномном институте Сингапура и представленные в GEO NCBI (http://www.ncbi.nlm.nih.gov/geo/)

по секвенированию методами ChIP-PET (GSE18046 ERα человека), ChIP-seq и ChIA-PET. Исходные данные секвенирования ChIP-seq представлены архивами GSE11431 для факторов Nanog, Oct4, STAT3, Smad1, Sox2, Zfx, c-Myc, n-Myc, Klf4, Esrrb, Tcfcp2l1, E2f1, CTCF и регуляторов транскрипции p300 and Suz12, GSE19219 для Tbx3, GSE17439 и GSE17642 для Eset, GSE19019 для Nr5a2, GSE23581 для Smad2 ЭСК мыши, GSE26831 и GSE23893 для ERα человека (культуры клеток MCF-7), GSE22767 и GSE22792 для PRDM14 (ЭСК человека) и прямое секвенирование ДНК (GSE26392 – нуклеосомная ДНК дрожжей). Данные ChIA-PET представлены архивами GSE18046 для ERα, и GSE33664 – для РНК-полимеразы II человека.

Использовались данные по структуре хроматина в масштабе полного генома, связанные с метилированием гистонов, составляющих структуру нуклеосомы (в частности гистона H3 – H3K4me3, H3K27me3, H3K4me1 и др.). Использовались также полногеномные данные по ацетилированию гистонов H3K9ac и H3K14ac, а также данные по доступности ДНК для белкового связывания, включая прямое секвенирование нуклеосомных фрагментов ДНК после обработки ультразвуком. Профили доступности ДНК в хроматине включали данные FAIRE, полученные с помощью выделения белковой фракции, связанной с ДНК, и последующего секвенирования [271]. Использовались данные по определению участков ферментативного разрезания ДНК с помощью DNAse I [316]. В настоящей Главе представлены компьютерные методы и программы обработки таких данных, разработанные автором.

2.2 Компьютерная обработка данных ChIP-seq

Для получения полногеномного распределения сайтов связывания различных транскрипционных факторов использовались данные ChIP-seq - метода иммунопреципитация хроматина с последующим секвенированием выделенных геномных фрагментов. Для каждого выделенного с помощью иммунопреципитации фрагмента ДНК, составляющего обычно от 150 до 300 нуклеотидов, секвенируется его часть – обычно первые 20-75 нуклеотидов (длина в зависимости от технологии), образуя прочтение ДНК, или «рид» (read). Для компьютерного анализа, представленного в настоящей работе, первичные данные секвенирования поступали в FASTA формате до картирования на геном, либо в формате bed-файлов, содержащих позиции секвенированных фрагментов в хромосомных координатах.

Такие геномные данные ChIP-seq, при расположении по хромосомам представляют собой профиль регуляторной активности (например, связывание ДНК с

белком, или со специфичными гистонами). Задача состоит в упорядочении и поиске функциональной информации, что требует как технических решений (хранение, поиск, быстрый доступ), так и новых теоретических подходов, связанных с математическим исследованием сигнала профиля ChIP-seq - одномерного численного профиля, привязанного к геномным координатам (позициям на хромосомах).

Как отмечалось в предыдущей Главе, дополнительная сложность состоит в том, что многие белковые факторы не работают по отдельности in vivo, формирую комплексы с другими ТФ и таким образом, могут связывать ДНК прямо или опосредованно [227] (так называемый эффект непрямого, опосредованного связывания ТФ - “piggy-back”).

Первым этапом обработки данных ChIP-seq является картирование ридов на геном - то есть определение для каждой последовательности хромосомы, позиции на хромосоме и ориентации на хромосоме (прямая или обратная). Использовались стандартные программы картирования для форматов Illumina/Solexa, инструмента Genome Analyzer - программы анализа изображений (флюоресценции меченых нуклеотидов на образце по технологии Illumina) Firecrest и Bustard, и программы анализа последовательностей Gerald и Eland.

Модуль GERALD (аббревиатура от «Generation of Recursive Analyses Linked by Dependency») предназначен для выравнивания секвенированных последовательностей и фильтрации данных по качеству. Модуль GERALD содержит программу ELAND для картирования прочтений на референсный геном. В нашей работе использовались данные прочтений ChIP-seq после картирования ELAND. Пример выравнивания прочтений после фильтрации ELAND по качеству показан на рисунке 2.1.

На языке C++ автором была написана программа анализа таких данных в текстовом формате и построения геномного профиля в формате bed-файла для заданного геномного релиза (релизы генома человека hg17, hg18, hg19, генома мыши mm8, mm9, с возможностью дальнейшего расширения на другое число хромосом и их размеры). Профиль далее анализировался либо собственными программами, либо стандартными средствами (программы MACS, CCAT).

Рисунок 2.1 содержит типовой пример данных секвенирования ChIP-seq: фрагменты ДНК из ChIP-seq эксперимента для Nanog после картирования на мыши, формат Illumina. Представлена нуклеотидная последовательность, цифры качества прочтения, координата положения фрагмента ДНК на хромосоме, ориентация на хромосоме (F или R, соответственно от «forward» или «reverse» - прямая или обратная) и нуклеотидная последовательность хромосомы (возможно 1-2 несовпадения).

На нижней панели рисунка показана схема расположения секвенированного прочтения (рида) ДНК, длиной от 25 нт, обозначенного черной стрелкой на последовательности всего выделенного после иммунопреципитации фрагмента (типично 150-300 нт), обозначенного серой стрелкой.

```
AAAAGCAACTTAGAGATTGCACCAC 12500 1 chr9:42004400 F AAAAGCAACTTAGAGATTGCACCAC 9359
GGTTTTATTATTATTTTAGGGGTTT 11453 1 chr13:116992396 F GGTTTTATTATTATTTTATGGGTTT 9359
GTAATGTGTTTTTTGTGACTATTTT 11453 1 chr12:115188071 R AAAATAGACACAAAAAACACATTAC 9359
AAGATGCCAGAACTCTCAGCTCCTT 12500 1 chr7:135122171 F AAGATGCCAGAACTCTCAGCTCCTT 11453
TGCAAGCTTTCTTATTCCCTTTCAT 10406 1 chr6:128978691 R ATGTAAGGGAATAAGAAAGCTGGCA 9359
...
```

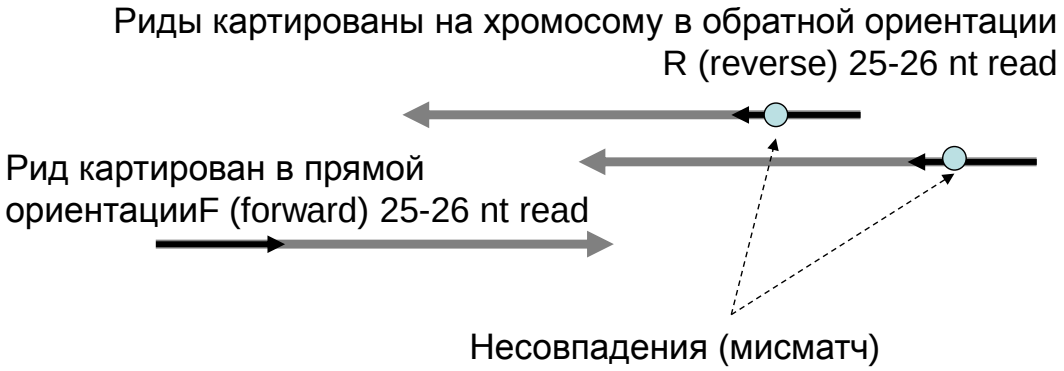


Рис. 2.1. (Верхняя панель) Фрагмент данных секвенирования ChIP-seq, формат Illumina (фрагменты ДНК из ChIP-seq эксперимента для Nanog, после картирования на мыши). Представлена нуклеотидная последовательность, цифры качества прочтения, координата положения фрагмента ДНК на хромосоме, ориентация на хромосоме (F или R, соответственно прямая или обратная) и нуклеотидная последовательность хромосомы (возможно 1-2 несовпадения).
(Нижняя панель) Схема расположения секвенированного рида (короткого фрагмента ДНК, длиной от 25 нт), обозначенного черной стрелкой на последовательности после иммунопреципитации (типично 150-300 нт), обозначено серой стрелкой. Ориентация положения в хромосомных координатах может быть прямой или обратной, возможны несовпадения с референсной последовательностью генома.

Такие данные обрабатывались написанной автором компьютерной программой для перевода файлов в bed-файл.

Распределение ошибок прочтения последовательности ДНК с помощью технологии Illumina неравномерно и имеет сдвиг к 3’-концу (последние 5-10% от длины прочтения), поэтому для более надежного определения последовательности возможно использование более короткой последовательности, а также дополнительная фильтрация по качеству [289]. Заметим, что в процессе развития технологии с 2006 года увеличилась длина прочтения Illumina (с 35 до 100 нт) и, в настоящее время, техническая проблема качества коротких прочтений отходит на задний план. В данной

работе выполнялся компьютерный анализ данных секвенирования ChIP-seq после картирования, с известными координатами на хромосомах референсного генома.

2.2.1. Компьютерный анализ профиля связывания ChIP-seq в геноме и статистическое определение пиков

На рисунке 2.2 представлена схема компьютерного определения кластеров фрагментов ДНК в хромосомных координатах. При построении профиля связывания координаты коротких (от 20 нт) прочтений удлиняются на размер фрагмента ДНК (150-200 нт) с помощью разработанной автором компьютерной программы на языке C++. Получается «ступенчатая лестница» (профиль) фрагментов, наложенных друг на друга, в линейных геномных координатах. Первоначальная компьютерная задача состоит в определении пиков такого профиля. Пик – наиболее высокая точка профиля, содержащая наибольшее число пересекающихся фрагментов. Пик наиболее вероятно содержит сайт связывания транскрипционного фактора, который физически связался с исследуемыми фрагментами в ChIP-эксперименте (рис. 2.2) [3, 9].

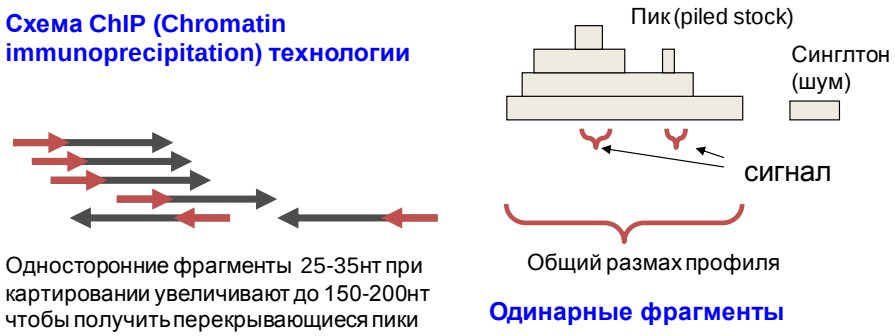


Рис. 2.2. Схема построения пика профиля ChIP-seq.

Рассмотрим разметку, полученную в результате ChIP-эксперимента по секвенированию специфически связанных фрагментов ДНК (рис. 2.2). Пик можно формально описать набором параметров, включая вершину («саммит»), начало и конец (общий размах). Пик может иметь несколько вершин (мультимодальный пик). Схема построения работает для парных фрагментов (парных концов, полученных с помощью метода ChIP-PET) и одиночных фрагментов ChIP-seq.

Для исследованных факторов транскрипции в геноме мыши и в геноме человека было получено от 1000 до 30000 пиков, соответствующих сайтам связывания транскрипционных факторов [3, 13]. Обычное значение для эксперимента ChIP-seq – около 5 тысяч сайтов в геноме, каждое со своей высотой пика, характеризующей его «силу». Высоту пика можно трактовать как сродство белка к ДНК при иммунопреципитации. Эффект подтверждается выборочным независимым

тестированием, содержащихся в выделенных пиках профиля ChIP-seq, по параметрам амплификации связанных с исследуемым белком нуклеотидных последовательностей с помощью количественной ПЦР (кПЦР). Показано, что высота пика в полногеномном эксперименте ChIP-seq коррелирует с силой связывания в отдельном биохимическом эксперименте кПЦР [13].

Поиск мотивов сайтов связывания

Анализ нуклеотидных последовательностей, соответствующих пикам профиля ChIP-seq в геноме, позволяет найти нуклеотидные мотивы связывания исследуемых ТФ. Так, на рисунке 2.3 показан пример расположения консенсуса CACGTG (E-бокс) для связывания транскрипционного фактора MYC в нуклеотидной последовательности в локусе гена человека NPM1, содержащейся в пике профиля, построенного по методу ChIP-PET [9].

Уточнение мотивов сайтов связывания в выделенных нуклеотидных последовательностях, соответствующих пикам профилей ChIP-seq, далее выполняется с использованием компьютерных программ MEME (http://meme.nbcr.net/meme/) и Weeder [230]. Для анализа контекста нуклеотидных последовательностей сайтов связывания транскрипционных факторов также использовались алгоритмы анализа частот нуклеотидных l-грамм [58], и разработанные ранее оценки сложности генетических текстов [57, 114].

В результате шума сигнала профиля ChIP-seq (ошибки секвенирования) и неспецифического связывания в геномном профиле могут получаться ложные пики (там, где был другой белок, или была неверно картирована другая последовательность ДНК). Задача анализа ChIP-seq состоит в том, чтобы отличить истинные пики от ложных, неспецифических пиков. Истинные пики, как правило, значительно выше, что можно показать статистически. Истинные пики ChIP-seq чаще содержат мотив сайта связывания исследуемого в эксперименте транскрипционного фактора. Известные нуклеотидные мотивы (весовые матрицы) связывания транскрипционных факторов представлены в описанных в Главе 1 базах данных, таких как JASPAR, TRANSFAC, TRRD [121], как результат компиляции разрозненных экспериментов. Высокий процент последовательностей пиков ChIP-seq, содержащих известный для исследуемого белка мотив связывания, служит подтверждением корректности эксперимента и правильности выделения пиков геномного профиля.

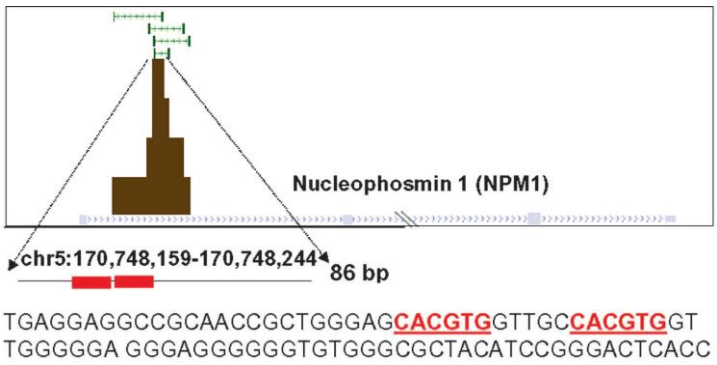


Рис. 2.3. Поиск известных мотивов связывания транскрипционных факторов в нуклеотидных последовательностях профиля ChIP-seq (на примере ТФ c-Myc) [9].

Встает задача – правильно ранжировать пики используя статистические оценки вероятности получения такого пика по случайным причинам. Используется контрольное секвенирование, дающее профиль распределения прочтений ДНК по хромосомам, полученный в результате эксперимента в тех же условиях, но без специфических антител, «пустой» прогон секвенатора. Поскольку контрольное секвенирование обычно не покрывает весь геном, необходимо определить статистически, сколько коротких фрагментов случайным образом попадет на данный участок ДНК в районе исследуемого пика (1000 нт и более). Полная картина требует оценки такой вероятности расположения прочтений для каждого короткого геномного интервала на хромосоме (размером вплоть до нуклеотида), что выполняется с помощью многократной компьютерной симуляции расположения прочтений, используя генератор случайных чисел (симуляция по методу Монте-Карло).

Таким образом, кроме высоты пика, есть параметр его значимости, вероятность получения по случайным причинам, по отношению к контролю (случайному распределению) - ошибка перепредсказания, или FDR (False Discovery Rate) (например, 5%). Дополнительным требованием является превышение высоты пика над высотой профиля в локальном окружении в геноме (обычно в 3-5 раз, как было предложено в [3]).

Задача определения набора пиков в геноме решается статистически с помощью сравнения экспериментально полученного распределения набора пиков к ожидаемому по случайным причинам. Автором предложен компьютерный алгоритм определения пиков профиля, 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 [16, 38].

Рассмотрим статистический подход по выделению пиков на примере транскрипционного фактора Nanog в геноме мыши. График на рисунке 2.4 показывает

наблюдаемое распределение кластеров ChIP-seq (высоты пиков профиля) и средний размер таких пиков в компьютерном эксперименте.

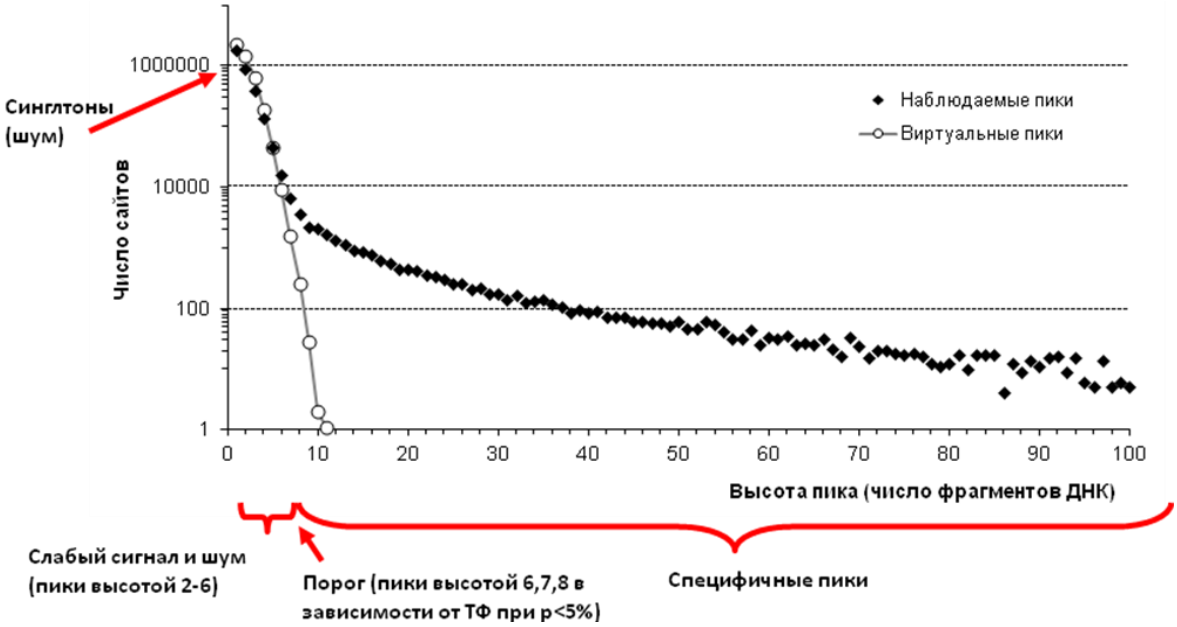


Рис. 2.4. Распределение числа пиков профиля ChIP-seq заданной высоты (сайтов) в геноме в зависимости от высоты пика (на примере профиля связывания ТФ Nanog в геноме мыши) [3, 38].

Такой подход основан на использовании наблюдаемых и контрольных данных связывания хроматина и компьютерном фильтровании шума в сигнале ChIP связывания (ложных пиков). Для статистического подтверждения корректности определения пиков использовали компьютерную симуляцию случайного сигнала связывания посредством генерирования распределения в геномных координатах того же числа виртуальных последовательностей.

Данный компьютерный подход применялся как для анализа данных ChIP-PET и картирования сайтов связывания транскрипционных факторов p53, MYC, STAT1 [9], так и для анализа данных ChIP-seq для факторов ER, FOXA1, PRDM14 в геноме человека [13, 42]. Геномные карты сайтов связывания опубликованы в представленных статьях [13, 42]. Тот же подход использовался для анализа данных ChIP-seq и определения сайтов связывания транскрипционных факторов Oct4, Nanog, Sox2, n-Myc, c-Myc и ряда других факторов в работе [3], факторов Tbx3 [41], Eset [39], Nr5a2 [40], Smad2 [54] в геноме мыши. Был выполнен анализ данных ChIP-seq и построена карта сайтов связывания транскрипционного фактора Zic3 в геноме рыбы *Danio rerio* [43].

2.2.2. Определение статистической значимости найденных пиков профиля связывания ChIP-seq

Число фрагментов ДНК в ChIP-эксперименте покрывающих специфические сайты связывания на хромосомах должно в целом соотноситься с аффинностью (сродством) сайта связывания к белку - фактору транскрипции. Действительно, чем выше аффинность, тем больше шанс для белка быть связанным в исследуемом пуле клеток, и тем выше вероятность получить несколько связанных фрагментов ДНК из одного геномного локуса при секвенировании. В то же время какие-то фрагменты ДНК быть получены случайно или ошибочно картированы в геноме. Для наблюдаемого распределения вероятности получения пиков заданной высоты в ChIP-эксперименте была предложена модель взвешенной суммы (смеси) распределений [16], включающая специфичные сайты связывания (истинный сигнал) и неспецифичные секвенированные фрагменты ДНК (шумовой сигнал).

$$P_{ob}(X=m) = \alpha * P_{sp}(X=m) + (1-\alpha) * P_{ns}(X=m), \quad (2.1)$$

здесь P_{ob} распределение вероятности встретить пики ChIP профиля высоты m, X - высота пика (размер кластера пересекающихся последовательностей), m=1,2,3,... число фрагментов ДНК, составляющих пик, P_{sp} - распределение вероятности встретить специфические пики ChIP профиля такой высоты, $0<\alpha<1$ - доля специфичных (связанных с исследуемым белком) последовательностей в ChIP эксперименте, P_{ns} - распределение вероятности неспецифичных (шумовых) пиков. Параметр α соответствует «весу» специфичного распределения в смеси и должен быть связан с качеством антител для иммунопреципитации белка.

Используя такой статистический подход смеси истинного и шумового сигнала, первоначально примененный к ChIP-PET экспериментам, можно моделировать и другие параметры пиков профиля ChIP-данных - общее число перекрывающихся фрагментов ДНК в кластере, закрывающим специфический сайт связывания (высота пика), размах кластера таких фрагментов в геномных координатах [9]. Специфическое распределение высот пиков P_{sp} может быть оценено обобщенной функцией Парето - линейно убывающей в логарифмических координатах прямой [16]. Чем больше высота пика, тем меньше вероятность его наблюдения, и эта вероятность стремительно убывает обратно пропорционально высоте пика. Такая форма распределения математически связана с распределением аффинности связывания ДНК с белком [16].

«Шумовое» распределение P_{ns} может быть построено с помощью компьютерных симуляций, предполагая равномерное распределение неспецифичных последовательностей вдоль хромосом. Пример распределения вероятностей

наблюдаемой и ожидаемой высот профиля ChIP-PET (PET-кластеров перекрывающихся фрагментов ДНК) для эксперимента по связыванию белка p53 приведен в Таблице 2.1.

Таблица 2.1

Число ChIP-PET прочтений ДНК, образующиеся кластеры прочтений (пики профиля) и оценки случайного образования пиков

	Всего PET	Высота пика (число прочтений)							
		Одиночные PET	PET-2	PET-3	PET-4	PET-5	PET-6	PET-7	PET-8+
Оценка числа пиков с помощью симуляции Монте-Карло	65,714	64,943	770.9	11.1	0.0034	0*	*	*	*
Наблюдаемое число пиков PET		61,270	1,443	160	66	38	29	13	27

* - не вычислимо в явном виде, меньше 0.0001

С помощью разработанной автором компьютерной программы на языке C++ генерировалось случайное распределение позиций в геноме (по размеру хромосом мыши в референсном геноме) для точно такого же числа прочтений, как и в эксперименте ChIP-seq. В программе использовался датчик случайных чисел с большим периодом, чтобы избежать искусственной кластеризации. Каждая позиция удлинялась на 200 нт, что соответствовало размеру фрагментов ДНК в эксперименте, и строился полногеномный профиль. Применялся алгоритм выделения ChIP-seq пиков, подсчитывалось число и размер пиков в геноме.

Таким образом, использовались два профиля - один для наблюдаемого реального ChIP-seq связывания, второй для профиля, ожидаемого по случайным причинам в линейных геномных координатах.

Компьютерное симулирование равномерного распределения фрагментов по хромосомам может быть упрощенной моделью, поскольку некоторые участки хромосом (прежде всего гентерохроматин), в целом менее доступны для секвенирования. Поэтому по возможности в качестве контроля должен быть взят экспериментальный профиль прямого секвенирования доступной ДНК в геноме. Такой контрольный профиль секвенирования может быть взят и из эксперимента ChIP-seq с неспецифическим антителом - белком, который не связывается с ДНК - например IgG (иммуноглобулином G) или флюоресцирующим белком GFP (green fluorescent protein). Такой контрольный профиль для GFP использовался в работе автора [3].

Выделение неслучайных пиков профиля для заданного транскрипционного фактора выполняется в два этапа. Во-первых, определяется минимальная высота пика,

при которой его можно считать неслучайным. Высота пика определяется как число секвенированных фрагментов ДНК, которые пересекаются между собой на геномной карте, находятся в одном геномном локусе. Пороговое значение высоты пика определяется через оценку ошибки перепредсказания сайта связывания в пике данной высоты, т.е. вероятность (уровень) ложного предсказания FDR (false discovery rate, в англоязычной литературе). Чем меньше ошибка, тем лучше предсказание. Использовался FDR в 1%. Высота пика, при которой число пиков в случайном (контрольном распределении) не более 1% пиков от распределения числа пиков в ChIP-seq эксперименте, использовалась как пороговое значение (обычно это высота от 3 до 6).

Во-вторых, для каждого пика индивидуально определяется отношение его высоты к высоте профиля контрольного неспецифического секвенирования. Этот параметр называется «кратное увеличение» (fold-change). Такой подход нужен, чтобы отсеять высокие пики профиля, если в конкретном геномном локусе контрольное секвенирование также дает высокие пики.

Типичное распределение плотности прочтений ChIP-seq для специфичного и неспецифичного связывания вдоль хромосомы показано на рисунке 2.5 (для хромосомы 2 мыши).

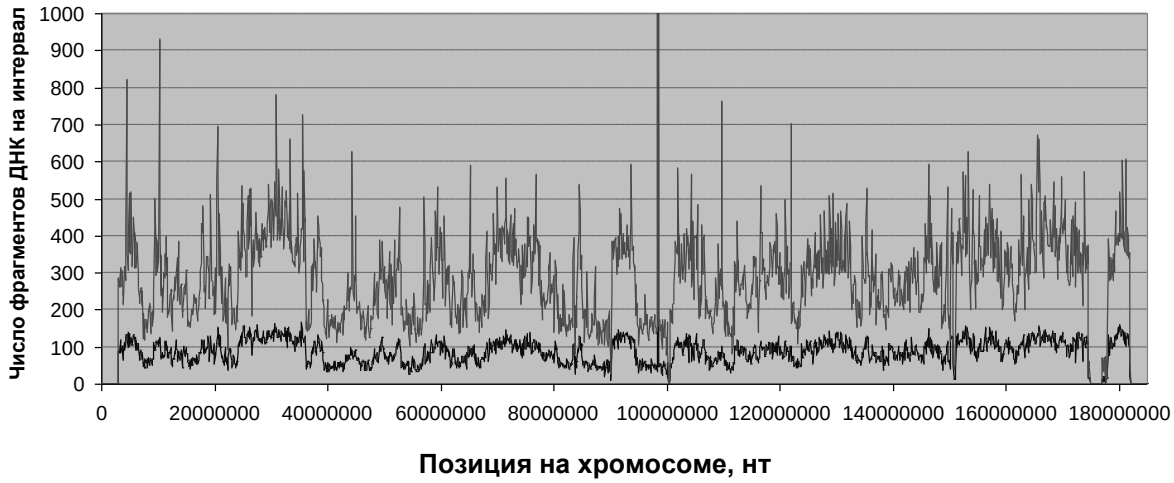


Рис. 2.5. Профили распределения ChIP фрагментов вдоль хромосомы для специфического (Nanog) и неспецифического связывания (GFP) на хромосоме 2 мыши. По оси ординат - число прочтений ДНК (суммарно по нуклеотидам) в интервале 100 нт. Вверху профиль связывания транскрипционного фактора Nanog, внизу - контрольный профиль связывания GFP.

Видно неравномерное распределение профиля контрольного секвенирования по хромосоме для неспецифического белка GFP (нижний профиль). Наблюдается сходная форма профилей - в масштабе хромосомы, на больших интервалах, профили

секвенирования коррелируют между собой. Однако такая корреляция обусловлена шумовым сигналом, неспецифическим связываниям и технологическими причинами.

Видны высокие пики связывания транскрипционного фактора Nanog (верхний профиль) и отсутствие выраженных острых пиков для профиля контрольного секвенирования (нижний профиль). В то же время наблюдается сходная форма профилей (корреляция). Для профиля контрольного секвенирования нет равномерности распределения прочтений вдоль хромосомы.

После дополнительной фильтрации пиков ChIP-seq относительно контрольного секвенирования по отношению высот получаем меньшее число пиков, но форма распределения пиков по высоте остается той же. Пример для того же транскрипционного фактора Nanog приведен на рис. 2.6.

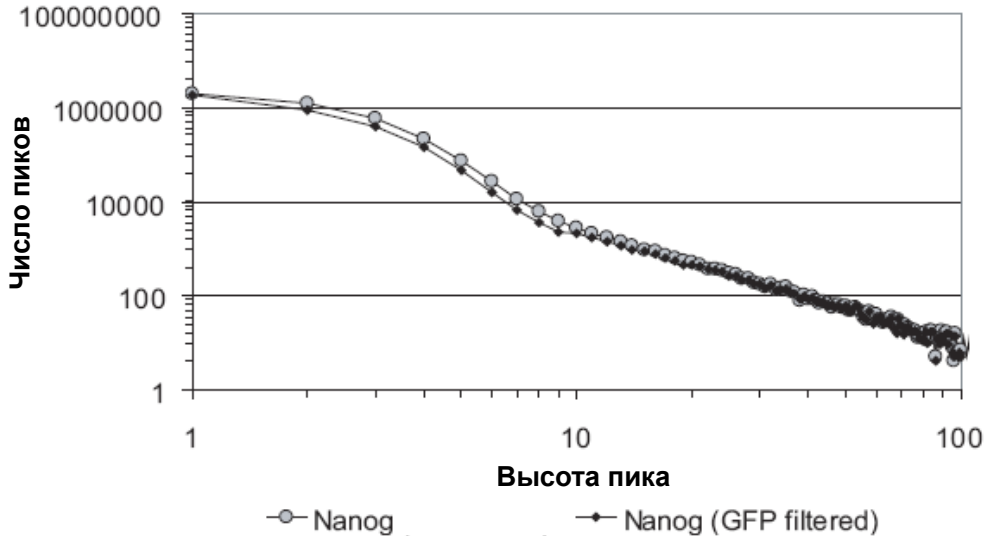


Рис. 2.6. Распределение высоты пика для ChIP-seq библиотеки ТФ Nanog до и после фильтрации относительно контрольной библиотеки секвенирования (GFP) в логарифмической шкале (log-log) [3]. По оси абсцисс – высота пика, по оси ординат – общее число пиков профиля ChIP-seq.

Причинами неравномерности распределения прочтений на хромосоме могут быть как смещение GC состава, большая доступность ДНК, свободной от белковой фракции при подготовке материалов эксперимента, а также чисто технологические причины, связанные с типом секвенирования (технологии 454, Illumina Solexa).

На первом этапе обработки данных ChIP-seq выполняется компьютерная симуляция Монте-Карло - случайное определение позиций в геноме для виртуальных фрагментов ДНК длиной 200 нт, предполагая равномерность их распределения между хромосомами и вдоль каждой хромосомы. Так, в работе [3], оценивалось распределение случайных фрагментов для генома мыши, использовался геномный релиз mm8. Пороговое значение высоты пика профиля связывание выбиралось так, чтобы

отношение числа неспецифических пиков данной высоты и выше к числу специфических пиков не превышало 0.05. Это число и является оценкой перепредсказания - т.е., ответом на вопрос, какова максимальная доля ложных пиков, выбранных из общего распределения только по высоте пика. В выполненных исследованиях использовался более строгий допустимый порог ошибки, на уровне 0.01 и ниже.

Общие статистические свойства сгенерированного на компьютере распределения пиков и числа пиков в контрольном секвенировании схожи: подавляющее большинство невысоких пиков, малое число высоких пиков, распределение пиков по высоте всегда быстро убывает, нет «выбросов» распределения для какой-то одной высоты. Модель равномерного распределения неспецифических фрагментов секвенирования в геноме имеет свои недостатки.

Рассматривая профиль контрольного секвенирования вдоль хромосомы (рисунок 2.5) легко видеть, что распределение по позициям хромосом уже не равномерно, заметно повышение числа фрагментов с чуть большим GC составом, присутствием сателлитных повторов (альфа-сателлиты и прицентромерные участки).

Таким образом, требуется второй шаг фильтрации ложных пиков посредством сравнения с негативным, или неспецифическим контрольным профилем секвенирования вдоль хромосом. В работе [3] для каждой ChIP библиотеки интенсивность профиля была нормализована (умножена на соответствующее число) в зависимости от глубины секвенирования (общего числа секвенированных фрагментов ДНК). Полученный набор пиков был сравнен с пиками в тех же позициях для контрольной ChIP библиотеки в тех же клетках (ЭСК мыши), где в качестве антитела использовался белок GFP. Пики, превышающие по размеру контроль более чем в пять раз (≥ 5), были выделены как специфичные сайты связывания [3].

Формула для оценки значимости пика по отношению высот профилей включает псевдосчетчик (чтобы избежать деления на ноль - минимальное число прочтений в контроле) и хорошо работает для достаточно больших значений высоты пика, в экспериментах работы [3] это высота - 5-10.

Для преодоления вычислительных сложностей компьютерная симуляция прочтений в геноме выполнялась многократно, до 1000 с оценкой средних и максимальных значений формирования кластеров. Особое внимание было уделено качеству генерации случайных чисел. Использовался опубликованный алгоритм генерации псевдо-случайных чисел MT19937 «Mersenne Twister» [100].

2.2.3. Фильтрация профиля связывания ChIP-seq по геномной аннотации

Профиль неспецифического связывания на хромосомах может иметь неожиданно высокие пики. Их число, как правило, невелико. На рисунке 2.7 показан профиль распределения ChIP фрагментов вдоль хромосомы для специфического связывания (Nanog) на хромосоме 14 в ЭСК мыши [3].

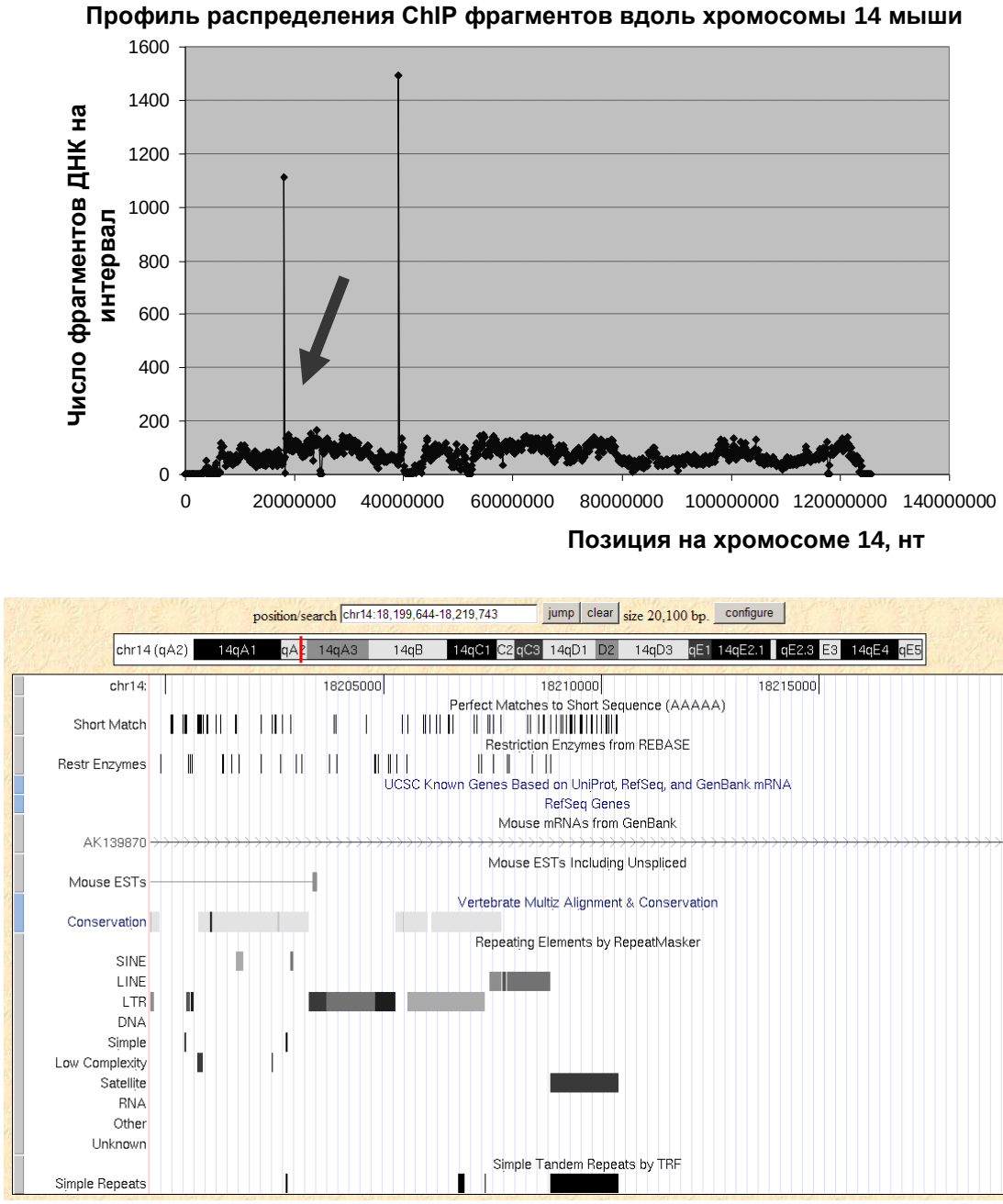


Рис. 2.7. (Верхняя панель) Профиль распределения ChIP фрагментов вдоль хромосомы для связывания Nanog на хромосоме 14 мыши. Стрелкой показан один из максимальных пиков. (Нижняя панель) Геномная аннотация участка первого высокого пика (позиция 18 200 000 на хромосоме 14 мыши) из UCSC Genome Browser. Видно наличие сателлитного повтора (простого тандемного повтора).

На рисунке по оси ординат - число прочтений в интервале 100 нт. Видны аномально высокие пики связывания Nanog (два пика в позициях около 18 млн. и около 40 млн. п.н.). Нижняя панель рисунка содержит геномную аннотацию участка первого высокого пика этого профиля (позиция 18 200 000 на хромосоме 14 мыши), представленную в UCSC Genome Browser. Видно отсутствие генов в данном районе генома и наличие сателлитного повтора (простого тандемного повтора), аннотированного RepeatMasker.

Таким образом, геномные повторы, являются источником ошибок при картировании прочтений. Поскольку повтор является простым (это не мобильный элемент, который потенциально может содержать сайты связывания транскрипционных факторов), речь идет именно о технической ошибке. При препроцессинге данных секвенирования такие пики, попадающие в сателлитные повторы, были удалены из дальнейшего анализа. Список сателлитных повторов из геномной аннотации RepeatMasker составил несколько десятков районов в бедных генами участках генома и не влиял на последующий анализ пиков ChIP-seq. Геномные координаты таких районов приведены в таблице в Приложении.

2.3. Метод оценки полноты (сатурации) эксперимента ChIP-seq

Оценка полноты (сатурации) эксперимента ChIP-seq по определению ССТФ в масштабе генома соответствует статистической оценке недопредсказания - оценке, сколько еще сайтов в геноме осталось не выявлено в данном конкретном эксперименте. Автором предложен оригинальный метод и алгоритм определения насыщенности (полноты) эксперимента ChIP-seq по определению сайтов связывания транскрипционного фактора в геноме с помощью пошаговых компьютерных симуляций и экстраполирования числа выявленных в геноме сайтов на каждом шаге. Метод опубликован в статьях [3, 38].

Рассмотрим проблему определения полноты (насыщения, сатурации) ChIP эксперимента и выявления всех специфичных сайтов связывания. Известно, что при увеличении общего объема, или глубины секвенирования (увеличении количества прочтений), выявляется больше сайтов, больше специфических участков. Глубокое секвенирование позволяет найти транскрипты, которые экспрессируются на очень низком уровне, и с трудом детектируются в обычном эксперименте. При увеличении объема данных ChIP-seq мы также можем найти новые, более слабые пики профиля, соответствующие слабому связыванию ТФ. В то же время увеличение объема данных, помимо дороговизны и трудоемкости эксперимента, приводит к увеличению шума и

ложных сигналов. Задача здесь – определить оптимум объема данных, при которых выявлены все - или почти все - специфичные сигналы сайтов связывания. Математически это означает, что нужно определить глубину секвенирования, при которой находятся 95% или 99% от общего числа сайтов, которые могут быть определены из профиля ChIP-seq полученного по данной технологии. Такая задача может быть определена путем анализа исходных данных секвенирования и компьютерной экстраполяции для фиксированной технологии (фиксирована специфичность связывания - качество антител для иммунопреципитации и тип секвенирования - длина прочтений).

Задача это не только компьютерная, но и технологическая, и статистическая, и биологическая.

Технологическая задача: Для фиксированного эксперимента секвенирования, с заданной глубиной секвенирования, имеющимися исходными данными распределения ридов по хромосомам (координат ридов) нужно определить достаточно ли данных для выявления всех специфичных сайтов – решить задачу определения полноты эксперимента.

Статистическая задача: определить ошибку недопредсказания сайтов связывания ТФ в геноме на имеющихся данных ChIP-seq от общего объема данных, который может быть определен с помощью такой ChIP технологии.

Биологическая задача: определить, сколько всего сайтов связывания данного транскрипционного фактора присутствует в геноме, насколько это определимо в уже проведенном ChIP эксперименте.

Имеющиеся данные для анализа – хромосомные координаты прочтений ДНК. Имеющиеся средства – процедура (компьютерная программа) для построения профиля связывания и выделения значимых пиков, соответствующих сайтам связывания. Задача решается с помощью компьютерной симуляции и обратной экстраполяции. Мы убираем (случайным образом) часть прочтений из расчетов, фиксируя некоторый меньший размер библиотеки секвенирования, пересчитываем профиль и число сайтов, которые определяются при таком уменьшенном объеме данных. Затем повторяем процедуру еще и еще, определяя, сколько ССТФ в геноме будет находиться при все меньшем объеме данных, вплоть до минимальных значений библиотеки секвенирования. Полученные данные формируют таблицу – объем библиотеки – число сайтов в геноме. Экстраполируя вперед можно определить функцию роста числа сайтов в зависимости от объема секвенирования и оценить число сайтов, которые могут быть получены при увеличении глубины секвенирования.

Использовались статистические оценки определения степени сатурации в ChIP-seq библиотеках на основе наблюдаемого распределения выделенных сайтов связывания. Рассмотрим в качестве примера ChIP-seq библиотеку для транскрипционного фактора Nanog в ЭСК мыши [3]. Для Nanog было получено 8,805,428 уникально картированных последовательностей (на геном мыши, релиз mm8). После удаления прочтений, ошибочно картированных на хромосомы Y и M (ошибочно, поскольку использовалась клетки от самок мыши, и только ядерный геном без митохондриальной ДНК), было получено 8,801,927 (99.96%) прочтений. После фильтрации «шумовых» кластеров прочтений, перекрывающихся с контрольной библиотекой GFP, было получено 6,684,737 прочтений (около 24% было отфильтровано).

Распределение получившихся перекрывающихся кластеров фрагментов ДНК по высоте пика и общему числу входящих в них последовательностей показано на рисунке 2.4. Наибольший кластер содержит 5,242 перекрывающихся ChIP фрагмента. Правая часть графика показано не полностью для лучшей визуализации основной части распределения. Для разделения специфичных и неспецифичных кластеров (пиков профиля) использовалась компьютерная симуляция формирования случайных кластеров при равномерном распределении прочтений ДНК по хромосомам. Случайные кластеры практически отсутствуют от 10 последовательностей формирующих кластер (высоты пика) и выше. Такая компьютерная симуляция дает минимальную оценку порогового значения высоты пика (числа последовательностей) в специфичных кластерах.

Минимальный размер кластера пересекающихся фрагментов ДНК может быть в дальнейшем уточнен экспериментально путем сравнения с контрольной библиотекой секвенирования и тестированием некоторого ограниченного набора последовательностей ДНК из этих кластеров в независимых экспериментах количественной ПЦР (qPCR). Так, для ChIP-seq библиотеки Nanog было определено пороговое значение высоты пика - 11 и число специфичных сайтов в геноме - 10343. Это пороговое значение было подтверждено выборочной проверкой нескольких сайтов посредством ChIP-qPCR (количественной ПЦР фрагментов ДНК после иммунопреципитации).

Рисунок 2.8 содержит пример расчета суммарного (кумулятивного) распределения пиков профиля ChIP-seq и общего числа фрагментов ДНК составляющих эти пики в зависимости от высоты пика ChIP-seq для связывания ТФ Nanog в ЭСК мыши.

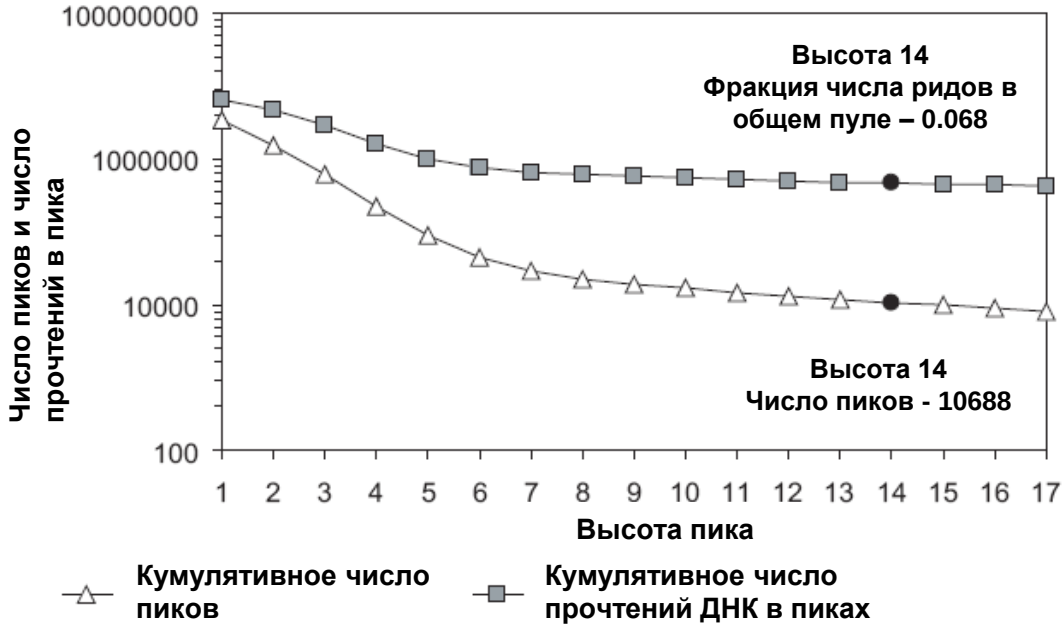


Рис. 2.8. Кумулятивное распределение пиков профиля ChIP-seq (помечено треугольниками) и общего числа фрагментов ДНК составляющих эти пики (помечено квадратами) в зависимости от высоты пика ChIP-seq (для связывания ТФ Nanog в ЭСК мыши) [3, 38]. Пороговое значение высоты специфичных пиков, подтвержденное выборочным тестированием qPCR, число пиков и общее число фрагментов ДНК в них отмечены черными кружками.

Далее мы определяем число фрагментов ДНК, входящих в эти пики профиля, и отношение этого числа к общему числу секвенированных фрагментов в данном эксперименте. Определяется число прочтений ДНК, формирующих специфичные кластеры (пики профиля) и оценивается их число относительно общего числа фрагментов в выборке. Рисунок показывает распределение общего числа пиков профиля и числа содержащихся в них последовательностей ДНК относительно высоты пика (здесь – размера кластера перекрывающихся последовательностей) в логарифмической шкале для сайтов связывания Nanog.

Заметим, что в то время как число пиков быстро убывает, число содержащихся в них последовательностей убывает значительно медленнее. На рисунке показано число специфичных прочтений для порогового размера высоты пика 14: 10688 пиков в геноме имеют высоту 14 и выше. Общее число прочтений, содержащихся в этих пиках, составляет 0.068.

Насыщение (сатурация) эксперимента ChIP-seq по определению общего числа сайтов в геноме в процессе секвенирования – это достижение момента, когда новые сайты уже не детектируются. Сатурация для заданного эксперимента ChIP-seq может быть промоделирована статистически, используя параметр специфичности связывания

последовательностей при иммунопреципитации и оценки общего числа сайтов в геноме. Оценка ошибки недопредсказания общего числа сайтов может быть найдена через число специфичных пиков профиля для библиотеки ChIP-seq к общему числу сайтов связывания в геноме, которое может быть определено при максимальном (потенциально бесконечном) увеличении глубины секвенирования.

Базовым предположения является то, что доля (фракция) специфичных прочтений, т.е. связанных с ССТФ, к общему числу всех прочтений N_{total} одна и та же, и определяется только специфичностью антитела к исследуемому транскрипционному фактору. Действительно, этот параметр – аффинность антител к ТФ - не зависит от того сколько раз выполнено прочтение ДНК в секвенаторе, т.е. от общего размера библиотеки. Это число может быть рассчитано из распределения найденных сайтов в профиле связывания. Для каждого пика профиля (соответствующих хромосомных координат) определим соответствующие им прочтения – общее кумулятивное число N_{cum}. Фактически это целое число - сумма числа N_{peak} всех пиков, умноженных на высоту пика *Height(i).*

$$N_{cum} = \sum_{i=min;max} (N_{peak}(i) * Height(i)) \quad (2.2)$$

Соответственно, доля специфичных прочтений определяется как N_{cum} / N_{total}.

Например, для библиотеки ТФ Nanog такая доля специфичных прочтений приблизительно ~6.8% от общей массы всех прочтений (Рисунок 2.8). Используя уровень 6.8% последовательностей, как обогащение специфичными ChIP прочтениями, мы сделали компьютерную симуляцию назад, уменьшая размер библиотеки с шагом 250,000 прочтений начиная с имеющегося числа ~6.5 миллионов последовательностей (после фильтрации прочтений по качеству прочтения и фильтрования контрольного неспецифического секвенирования GFP) назад до минимального размера (100 тысяч прочтений на геном). Убирали случайным образом (с помощью датчика случайных чисел) заданное число прочтений из текущей библиотеки, затем пересчитывали число и высоту пиков.

Из таких распределений мы можем определить минимальный размер пика, чтобы все такие и большие пики содержали кумулятивно 6.8% специфичных последовательностей. После этого мы вновь убрали случайным образом 250,000 прочтений из профиля и пересчитали распределение остающихся пиков по их высотам (симуляция назад). Каждый раз определялась минимальная высота пика для специфичных последовательностей. Минимальная высота пика для библиотеки Nanog варьировала от 3 (размер библиотеки 100,000) до 15 (реальный размер библиотеки).

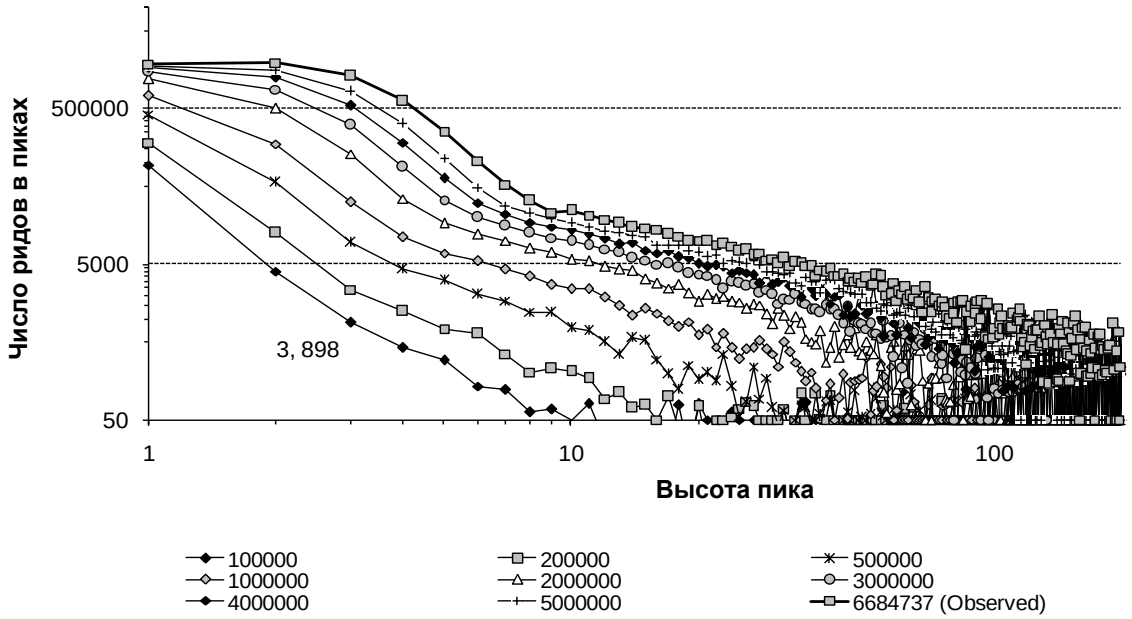


Рис. 2.9. Моделирование распределения числа пиков профиля ChIP-seq в геноме (ось ординат) в зависимости от высоты пика (ось абсцисс). Компьютерная симуляция назад для оценки зависимости числа предсказанных пиков от размера библиотеки. Показан пример для ChIP-seq библиотеки Nanog для генома мыши, Линии соответствуют уменьшению размера библиотеки (общего числа прочтений), распределенной по хромосомам [3, 38]. Исходное число картированных прочтений - 6684737. Далее из этого набора случайным образом удалялось заданное число прочтений (не меняя их позиций), и пересчитывались остающиеся пики.

На рисунке 2.9 представлено моделирование распределения числа пиков профиля ChIP-seq в геноме (ось ординат) в зависимости от высоты пика (ось абсцисс). Линии соответствуют уменьшению размера библиотеки (общего числа прочтений) [3, 38].

Исходное число картированных прочтений для ТФ Nanog - 6684737. Далее из этого набора случайным образом удалялось заданное число прочтений с шагом в 100 прочтений ридов, не меняя их позиций, соответственно менялся профиль распределения ридов на хромосомах и пересчитывались остающиеся пики. Заново строилась таблица числа ридов в библиотеке и распределение найденных пиков по высоте. Поскольку общее число пиков профиля уменьшалось, уменьшалось и пороговое значение числа найденных для данного профиля специфичных пиков (сайтов связывания).

На рисунке 2.10 представлена зависимость порогового значения высоты пика и числа выявленных пиков в геноме в зависимости от глубины секвенирования (для связывания ТФ Nanog мыши).

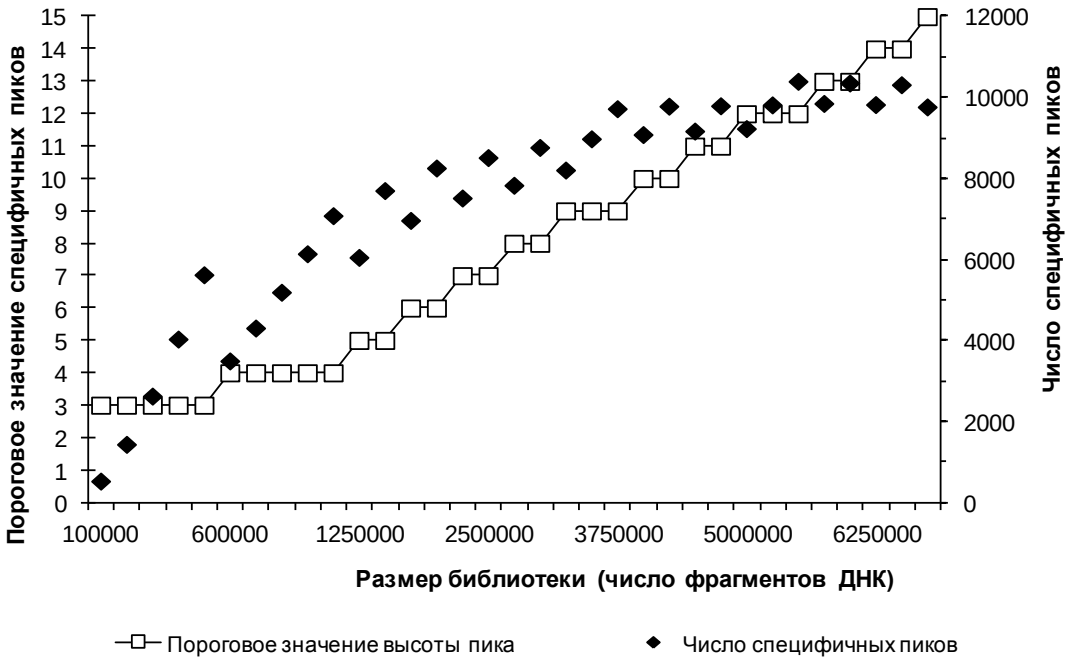


Рис. 2.10. Зависимость порогового значения высоты пика и числа выявленных пиков в геноме в зависимости от глубины секвенирования (для связывания ТФ Nanog мыши).

В то время как число специфичных последовательностей в библиотеке и соответствующее пороговое значение для специфичных пиков уменьшаются линейно, число специфичных сайтов уменьшается более медленно (см. кривые на рисунке 2.9 при движении справа налево – при уменьшении размера библиотеки). Можно предполагать, что при увеличении размера библиотеки (при движении слева направо и экстраполяции графика в будущее) закономерность роста сохранится.

Используя пороговое значение высоты пика, определенное по фиксированному проценту специфичных последовательностей, мы можем оценить число сайтов в геноме (специфичных пиков), которые могут быть найдены при различной глубине секвенирования. Поскольку пороговое значение высоты пика является целым числом, то число специфичных сайтов для фиксированного порога может быть чуть меньше или чуть больше чем заданное число специфичных фрагментов в библиотеке ChIP-seq. При компьютерных симуляциях профиля ChIP-seq по хромосомам используется датчик случайных чисел, и число получающихся пиков варьирует. Более правильно использовать минимальную и максимальную оценки числа пиков профиля (сайтов в геноме) при нескольких симуляциях. Использовались компьютерные оценки максимального (N') и минимального числа специфичных сайтов (N'') выделяемых при компьютерных симуляциях.

Далее выполнялась аппроксимация получившегося в результате пошаговой симуляции меняющегося размера библиотеки набора чисел аналитической функцией,

т.е. подбор параметров, с использованием пакета SigmaPlot 8.0. Была выбрана функция числа сайтов *f(X)* как функция размера библиотеки X, соответствующая формуле насыщения для связывания лиганда, применяемой в оценке параметров кинетики биохимических процессов. Функция определяет сатурацию эксперимента *f(X)* как пропорции числа сайтов в геноме, определялось по формуле:

$$f(X) = B_{max} * X/(K + X) \quad (2.3)$$

где B_{max} максимальное число сайтов в геноме (число насыщения эксперимента), а K – константа нормализации.

Таким образом, были найдены параметры аппроксимации, наиболее близкие рассчитанным данным (компьютерной симуляции кривой сатурации) для числа N" сайтов в геноме: B_{max} =12007.99, K = 21516.59

Таким же образом были подобраны параметры кривой, оценивающей минимальное число сайтов N'_1 и выполнена экстраполяция на больший размер библиотеки секвенирования (экстраполяция вперед).

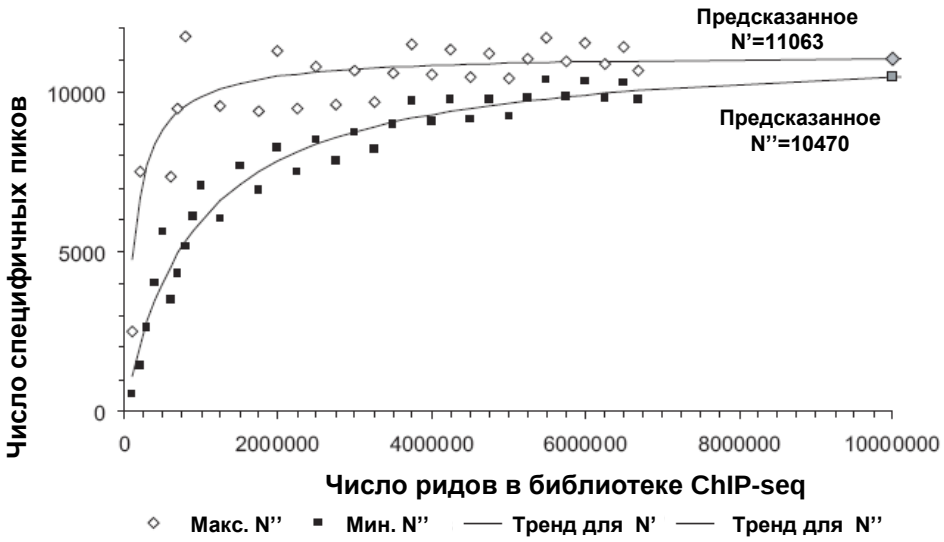


Рис. 2.11. Анализ сатурации (полноты) определения ССТФ в геноме в заданном эксперименте ChIP-seq [3, 38].

Минимальное и максимальное числа *N'* и *N''* специфичных пиков были оценены для уменьшенного размера библиотеки и экстраполированы на увеличенный размер (20 миллионов прочтений). Оценки этих чисел близко сходятся (разница меньше 5%), что дает число специфичных сайтов связывания Nanog в геноме. Сравнения параметров сатурации для фиксированного размера библиотеки (10 или 20 миллионов прочтений) позволяют оценить качество секвенирования для различных экспериментов иммунопреципитации.

Сатурация библиотеки ChIP-seq рассчитывалась как отношение оцененного

(экстраполированного) числа сайтов в максимально большой библиотеке (что эффективно соответствует числу сайтов в геноме) к числу сайтов в заданной фиксированной библиотеке (Таблица 2.2).

Таблица 2.2

Расчет числа сайтов и параметров экстраполяции для ChIP-seq библиотеки транскрипционного фактора Nanog

Размер ChIP-seq библиотеки Nanog	Пороговое значение высоты пика	Минимальное число сайтов N' при данном пороге	Максимальное число сайтов N'' (при пороге+1)	Аппроксимация минимального числа N'	Аппроксимация максимального числа N''
100000	3	2496	524	4770.8	1128.2
200000	3	7534	1432	6693.5	2053.6
300000	3	14477	2624	7732.3	2826.4
400000	3	22801	4033	8382.8	3481.5
500000	3	32708	5622	8828.4	4043.8
600000	4	7368	3493	9152.8	4531.8
700000	4	9508	4306	9399.4	4959.2
800000	4	11767	5190	9593.3	5336.8
900000	4	14280	6141	9749.8	5672.7
1000000	4	16954	7082	9878.6	5973.5
1250000	5	9589	6051	10119.4	6603.7
1500000	5	12438	7702	10286.6	7103.3
1750000	6	9431	6965	10409.4	7509.2
2000000	6	11293	8259	10503.4	7845.3
2250000	7	9509	7518	10577.7	8128.3
2500000	7	10827	8517	10638	8369.9
2750000	8	9615	7833	10687.7	8578.4
3000000	8	10685	8768	10729.6	8760.4
3250000	9	9692	8209	10765.3	8920.4
3500000	9	10588	8981	10796	9062.4
3750000	9	11506	9725	10822.8	9189.1
4000000	10	10564	9086	10846.4	9302.9
4250000	10	11359	9787	10867.3	9405.7
4500000	11	10490	9172	10885.9	9499
4750000	11	11210	9797	10902.6	9584.1
5000000	12	10434	9233	10917.7	9661.9
5250000	12	11065	9819	10931.4	9733.5
5500000	12	11694	10401	10943.8	9799.4
5750000	13	10983	9855	10955.3	9860.5
6000000	13	11544	10360	10965.7	9917.1
6250000	14	10887	9827	10975.4	9969.7
6500000	14	11413	10317	10984.3	10018.8
6684737	15	10688	9769	10990.5	10053

Для библиотеки Nanog это число составило 0.816 (81.6%). Таблица 2.2 представляет детальный расчет числа сайтов и параметров экстраполяции для ChIP-seq библиотек транскрипционного фактора Nanog.

Из таблицы 2.2 видно, что оценки максимального и минимального числа сайтов связывания транскрипционного фактора в геноме сходятся, более того, в аппроксимации даже пересекаются. Таким образом, моделирование позволяет получить устойчивую оценку числа специфичных сайтов в геноме, которые могут быть получены в ChIP-seq эксперименте в данных условиях.

Используя представленный выше алгоритм, было рассчитано число сайтов для 13 экспериментов ChIP-Seq в ЭСК мыши [3]. Общие оценки сатурации для всех ChIP-Seq библиотек в экспериментах для ЭСК мыши составили от 75 до 95%.

Следующая таблица представляет список исследованных транскрипционных факторов, число сайтов (пиков связывания ChIP-seq) в эксперимент и оценки полноты эксперимента (или сатурации), как доли сайтов связывания, корректно определенных в геноме по отношению к общему числу сайтов.

Таблица 2.3

Сатурация ChIP-seq библиотек транскрипционных факторов в ЭСК мыши

Фактор транс-крипции	Число сайтов в экспери-менте	Сатурация для секве-нирования 10 млн. прочтений	Сату-рация к об-щему числу	Сатурация (пределы оценки)	Максим. Число сайтов	Миним. Число сайтов	Оценка числа сайтов в геноме
c-Myc	3422	0.912	0.816	0.882-0.750	3881	3298	4400
CTCF	39609	0.932	0.910	0.942-0.879	40015	37366	42495
E2f1	20699	0.986	0.958	0.979-0.937	20750	19852	21190
Esrrb	21647	0.900	0.875	0.925-0.824	22895	20407	24754
Klf4	10875	0.822	0.751	0.825-0.677	10838	8894	13139
Nanog	10343	0.960	0.895	0.935-0.855	10688	9769	11426
n-Myc	7182	0.874	0.786	0.864-0.708	8258	6768	9555
Oct4	3761	0.869	0.784	0.886-0.682	3906	3007	4407
p300	524	0.891	0.762	0.801-0.723	556	502	694
Smad1	1126	0.799	0.747	0.890-0.604	1467	995	1648
Sox2	4526	0.876	0.757	0.831-0.683	5267	4331	6340
STAT3	2546	0.816	0.622	0.699-0.545	3031	2360	4334
Suz12	4215	0.901	0.802	0.871-0.734	4545	3827	5217
Tcfcp2l1	26910	0.978	0.929	0.955-0.902	26381	24900	27614
Zfx	10338	0.863	0.794	0.871-0.716	11465	9427	13157

Показана сатурация (полнота эксперимента) относительно секвенирования большей глубины (10 млн. прочтений), сатурация к общему числу сайтов, предельные

оценки максимального и минимального числа сайтов и итоговая оценка числа сайтов в геноме.

Как видно из таблицы, общий уровень сатурации эксперимента для ChIP-Seq библиотек в ЭСК мыши был достаточно высок, однако некоторые транскрипционные факторы, прежде всего STAT3, показали низкий уровень определения сайтов в геноме. Отметим, что представленная методика расчетов опирается только на геномные данные – карту расположения прочтений в ChIP-seq эксперименте. Использование другой культуры клеток и антител может изменить общие оценки числа сайтов.

На следующем рисунке приведена экстраполяция числа детектируемых сайтов в эксперименте ChIP-seq для ТФ STAT3 мыши.

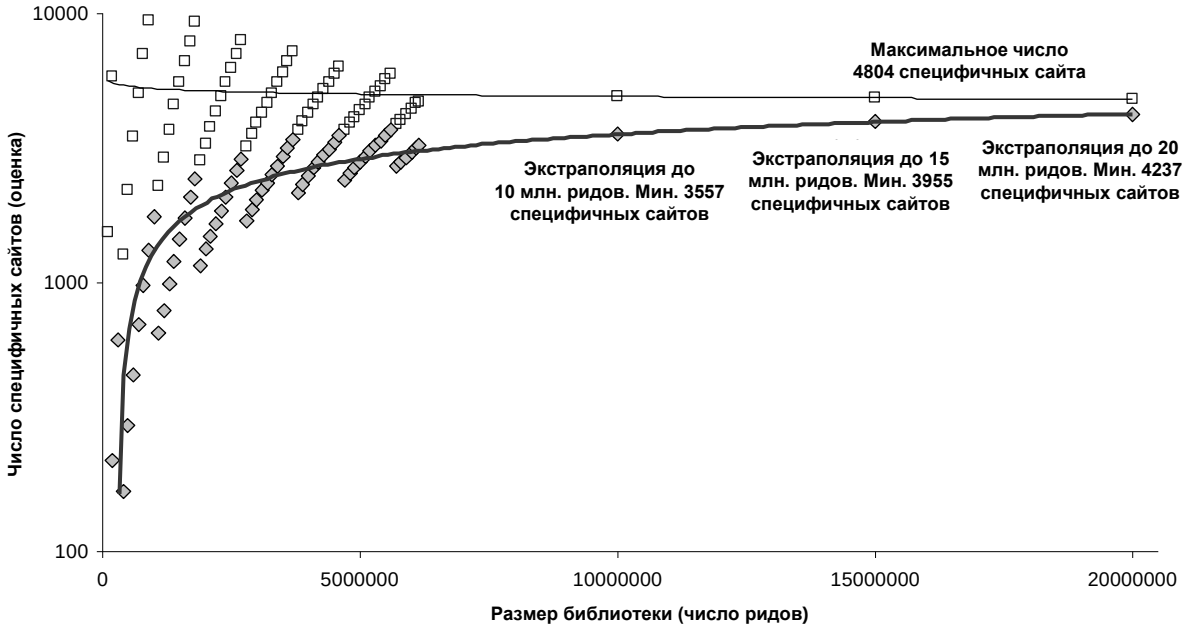


Рис. 2.12. Анализ сатурации (полноты) определения сайтов связывания в эксперименте ChIP-seq для транскрипционного фактора STAT3.

Отметим, что одним из выводов таких компьютерных расчетов становится утверждение о необходимости увеличения глубины секвенирования как минимум до 10 миллионов прочтений. Такая цифра сейчас, в 2012-2014 годах, является стандартом, в то же время на момент публикации данной работы [3] стандартом было 5-6 миллионов прочтений в библиотеке ChIP-seq.

2.4. Определение генов-мишеней транскрипционных факторов по данным экспрессии генов на микрочипах

Алгоритм определения генов-мишеней транскрипционных факторов по данным экспрессии генов на микрочипах основан на определении расстояния от ССТФ до старта транскрипции ближайшего гена, аннотированного в геноме (рис. 2.13).

Поскольку понятие «ближайший» может трактоваться неоднозначно в связи с расположением генов справа и слева от позиции на хромосоме, и существованием изоформ (несколько транскриптов начинаются с общего старта) была принята следующая процедура.

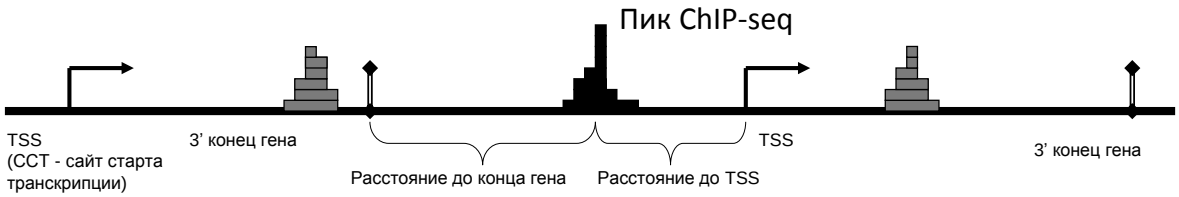


Рис. 2.13. Определение генов мишеней для ТФ по сайтам связывания в геноме.

Определяются координаты генов на хромосоме, выбираются минимальные по расстоянию. Положение сайта может быть в 5' районе или внутри гена, указываются оба параметра. Затем для списка позиций в геноме (сайтов или позиций профиля ChIP-seq), для каждого сайта составляется список координат генов-мишеней по расположению относительно этого сайта. Задавая интервал, например 10Кб, и перебирая все координаты генов, формируется список генов-мишеней. Такой подход был использован в работах [3, 9, 13, 37]. Разработана компьютерная программа оценки расположения сайтов (по списку, представленному только геномными координатами в bed-файле) относительно генов (по аннотации RefSeq генов в геноме), используя заданную классификацию расстояний и районов гена.

Рисунок 2.14 показывает пример такой классификации районов по вариантам расположения сайта относительно границ гена.

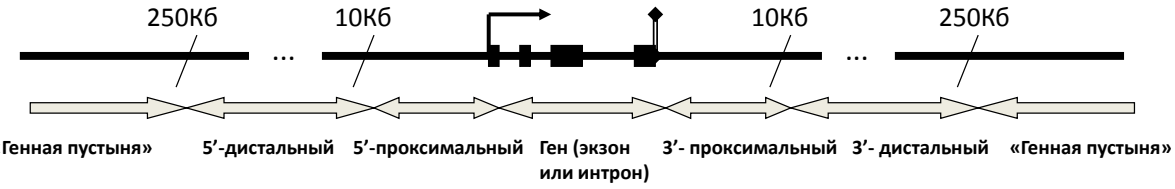


Рис. 2.14. Анализ распределения числа сайтов (пиков ChIP-seq) в зависимости от классификации положения относительно гена для сайтов c-Myc в геноме мыши.

Набор сайтов связывания может быть классифицирован по их геномной локализации, используя следующие определения:

- промотор (-5Кб до +1 Кб относительно старта транскрипции - TCC);
- внутригенные сайты: (+1 Кб от TCC до 3' конца гена);
- 3' район: (от 3' конца гена до 5 Кб после гена);
- 5' дистальные сайты: (от -100 Кб до -5 Кб относительно TCC);

- 3’ дистальные сайты: (от +5 Кб до +100 Кб поле 3’-конца гена) и «генная пустыня» (все остальные части генома).

Рисунок 2.15 представляет распределение сайтов связывания фактора c-Myc в геноме мыши относительно генов RefSeq.

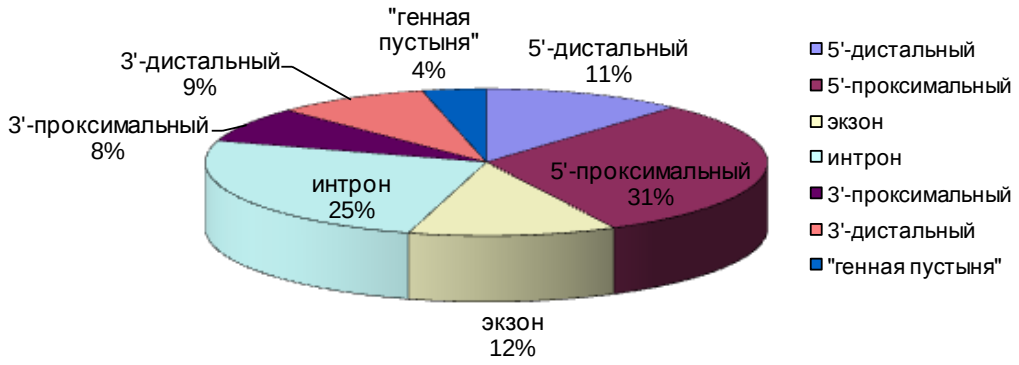


Рис. 2.15. Распределение сайтов связывания транскрипционного фактора c-Myc (пиков ChIP-seq) в геноме мыши относительно генов RefSeq.

Из рисунка 2.15 видно, что наибольшую долю составляют внутригенные (интронные) сайты, велика доля 3’- и 5’-дистальных районов. В то же время, доля сайтов в 5’-проксимальных к старту транскрипции районах (промоторах), где ожидается наибольшая концентрация регуляторных сайтов связывания, не является определяющей для всего полногеномного распределения сайтов.

Набор 16,043 сайтов связывания транскрипционного фактора ERα человека, полученных с помощью ChIP-seq, был классифицирован по геномной локализации, используя сходные определения, что приведены для предыдущего рисунка. Использовалась аннотация положения генов RefSeq уже в геноме человека, и тот же компьютерный алгоритм определения ближайшего гена, а затем классификации расположения сайта относительно его границ.

Расположение сайтов связывания транскрипционного фактора в промоторе другого гена позволяет рассматривать такой ген в качестве гена-мишени. Изменения экспрессии такого гена-мишени после активации транскрипционного фактора, детектированное с помощью микрочипов, подтверждает прямое действие транскрипционного фактора.

Рисунок 2.16 показывает появление пика связывания транскрипционного фактора ERα в геноме человека в окрестностях гена TFF1 после активации клеток линии MCF-7 эстрадиолом (правая панель). Видно, что экспрессия гена, являющегося мишенью связывания транскрипционного фактора - рецептора эстрогенов ERα, увеличивается в несколько раз.

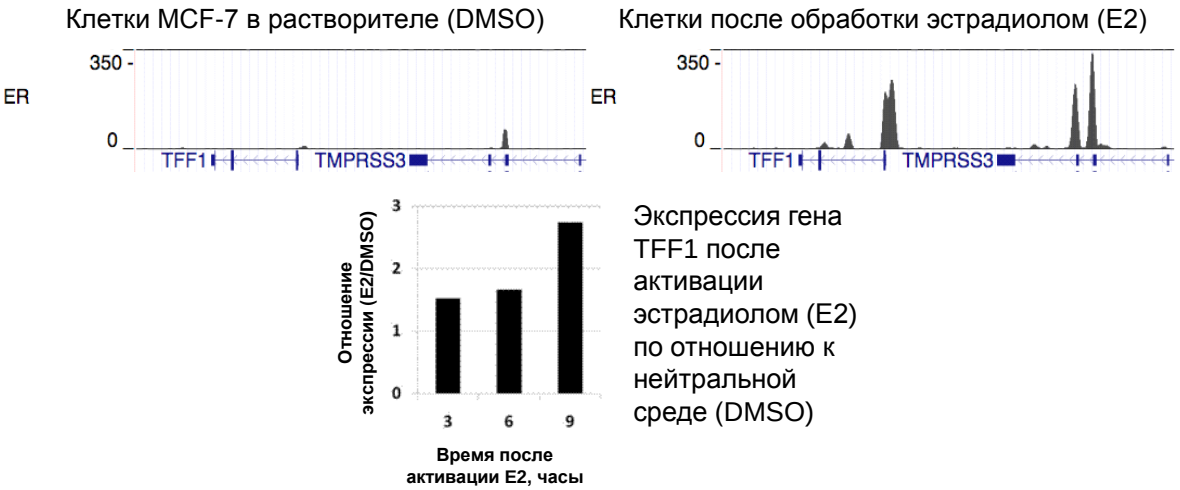


Рис. 2.16. (Верхняя панель) Профили связывания транскрипционного фактора ERα в промоторе гена TFF1 в геноме человека до и после активации эстрадиолом. (Нижняя панель) Увеличение уровня экспрессии гена TFF1 через 3, 6, 9 часов после активации ERα по данным микрочипов Affimetrix.

Действительно, высокая частота встречаемости сайтов связывания транскрипционного фактора в промоторах генов-мишеней подтверждается доступными данными по экспрессии генов на микрочипах.

Чтобы численно оценить значимость полученного распределения сайтов ERα в геноме, было сгенерировано контрольное распределение позиций в геноме, ожидаемое по случайным причинам. Использовались определения геномных районов относительно границ гена, приведенные выше. В качестве контроля было взято два распределения сайтов (рис. 2.17).

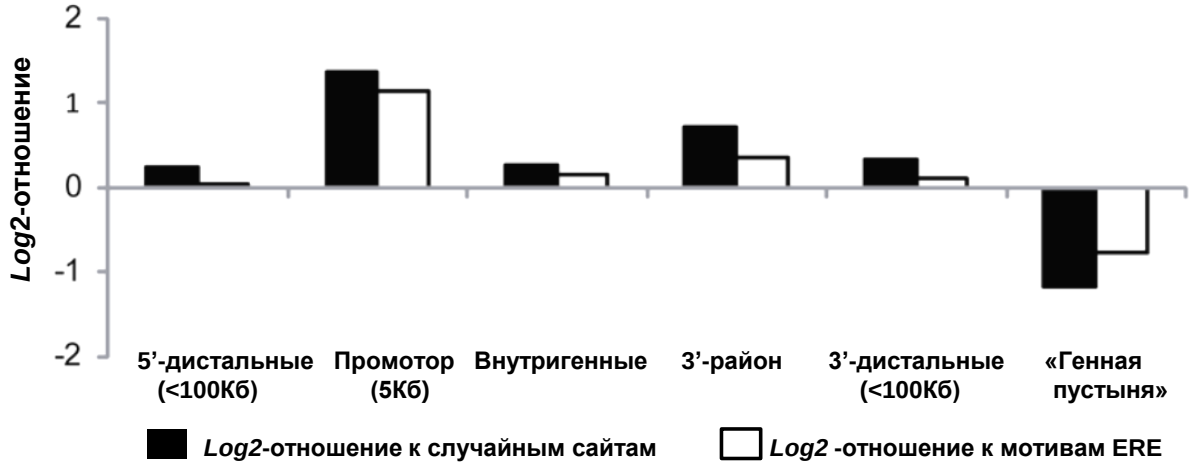


Рис. 2.17. Сравнение отношения числа сайтов связывания ERα, определенных с помощью ChIP-seq, по отношению к районам генов, к ожидаемому распределению случайных сайтов (белые столбики) и распределению мотивов ERE в геноме человека (черные столбики) в логарифмической шкале. Использовались те же геномные районы, что и в рисунке 2.16.

Первое - это распределение случайного набора сайтов в геноме человека (сгенерировано на компьютере с помощью датчика случайных относительно генов).

Второе контрольное распределение - это сайты, содержащие консенсусную последовательность связывания ERE (элемент ответа на эстрадиол) - 13 нт консенсус GGTCAnnnTGACC с не более чем одним несовпадением, размеченный собственной программой поиска гомологии в геноме человека. На графике показано отношение доли ChIP-seq сайтов ERα к случайным сайтам (черные столбики) и отношение той же доли к консенсусным сайтам ERE (белые столбики) в логарифмической шкале (log2).

Всего использовалось 56,746 сайта, содержащих ERE в геноме человека, причем эти сайты не были детектированы в пиках ChIP-seq. Из всего набора расположения участков генома, содержащих мотивы, были удалены те, которые были детектированы во всех опубликованных экспериментах ChIP-seq и ChIP-PET. Точные цифры статистической значимости (P-value) для отклонений полученных частот от ожидаемых по критерию Фишера показаны в Таблице 2.4.

Таблица 2.4.

Число сайтов и значения вероятности *P* для оценки различия наблюдаемого и ожидаемого числа сайтов связывания ERα в районах генов

	5' -дис-тальные	Промотор	Внутри-генные	3' конец гена	3'-дис-тальные	Удаленные («генная пустыня»)
Число сайтов						
ChIP-seq ERα сайты	2792	1485	6344	645	2299	2478
Случайные геномные участки	1473	359	3275	246	1141	3506
Сайты, содержащие ERE (GGTCAnnnTGACC) *	9675	2362	20297	1797	7603	15012
Вероятность различия						
ChIP-seq ERα сайты против случайных участков	1.32E-08	1.84E-73	1.56E-28	6.8E-12	9.33E-12	1.23E-286
ChIP-seq ERα сайты против консенсусных участков ERE	>0.1	4.42E-126	2.62E-18	2.05E-07	0.00248	4.29E-196

Прим.* - допускалось одно несовпадение в нуклеотидной последовательности с консенсусом ERE GGTCAnnnTGACC.

Геномные сайты, содержащие мотив ERE, но пересекающиеся с сайтами ERα, определяемыми ChIP-seq, были исключены из контрольной выборки. Всего использовалось 56,746 сайта с мотивом ERE в геноме человека (hg18).

Таблица содержит число сайтов связывания ERα из исследуемого набора и контрольные числа для сравнения. Показаны оценки уровня значимости (P-value) для отклонения наблюдаемого числа сайтов от ожидаемого, рассчитанные по точному критерию Фишера (2-сторонний критерий). Каждое число считалось по отдельности по таблице 2x2 (ожидаемое и наблюдаемое) для тестируемого района против всех

остальных районов генома. Показаны сравнения для случайного набора участков и для набора сайтов, предсказанных по консенсусу (с не более чем одним несовпадением, и не пересекающихся с известными экспериментальными данными).

Хотя наибольшая фракция сайтов ChIP-seq находится внутри генов (39%), а значительная часть в 5’-дистальных районах, эти районы не отличаются ни от ожидаемого по случайным причинам, ни от ожидаемого по нуклеотидной последовательности. В то же время, наиболее обогащены по сравнению с ожидаемым промоторные районы (приблизительно в два раза) и 3’-районы генов RefSeq. Фракция ChIP-seq сайтов ERα в «генной пустыне» значительно менее представлена даже по сравнению с сайтами, содержащими консенсус ERE, указывая на приближенность к генам реальных сайтов, найденных в ChIP-seq эксперименте. Косвенно это служит доказательством того, что компьютерное предсказание сайтов в геноме должно учитывать расстояние сайтов до генов. Около 15% сайтов, тем не менее, находится в «генной пустыне», и хотя это число меньше ожидаемого, должны быть молекулярные механизмы, связанные с дистальной регуляцией, утилизирующие использование удаленных сайтов связывания транскрипционного фактора.

2.5 Оценка качества сигнала экспрессии на микрочипах Affymetrix

Микрочиповый эксперимент является массовым экспериментом одновременной экспрессии тысяч генов в геноме, анализ которого необходим для исследований регуляции транскрипции. Важным этапом предобработки является проверка качества и фильтрация данных измерения экспрессии генов на микрочипах. Распространенная коммерческая платформа микрочипов Affymetrix имеет технические недостатки по аннотации генов [46]. Экспрессию мобильных элементов в опухолевых тканях была оценена статистически, по данным экспрессионных микрочиповых экспериментов.

Технология синтеза коротких олигонуклеотидных зондов (25 п.н.) непосредственно на поверхности микрочипа in situ с использованием литографических масок была разработана компанией «Аффиметрикс» для изготовления микрочипов GeneChip (Affymetrix, www.affymetrix.com/). Олигонуклеотидная матрица GeneChip использует наборы синтезированных in situ олигонуклеотидных проб, по 11–20 проб в наборе, каждая размером 25 нуклеотидов, для представления транскриптов генов или их изоформ. Для каждого гена-мишени использованы фрагменты-представители - целевые (таргетные) последовательности (initial target sequences) длиной 150–450 п.н. для выбора и локализации олигонуклеотидных проб. Дизайн проб (исходный выбор

производителем микрочипов локализации в гене и структуры олигонуклеотидных проб) может не соответствовать целевому транскрипту (гену-мишени) и содержать ряд технических проблем, связанных как с гибридизацией, так и с аннотацией – неверное указание гена-мишени, неоднозначность соответствия один набор проб–один ген [151, 156-159].

Анализ последовательностей олигонуклеотидных проб на микрочипах

Использовались данные о целевых нуклеотидных последовательностях (мишенях) для наборов проб Affymetrix, микрочипы серий U133A и U133B, загруженные с официального сайта разработчиков платформы NetAffx (http://www.affymetrix.com/analysis/index.affx) [148]. Эти последовательности предназначались для однозначной детекции транскрибируемых последовательностей в геноме, что требовало независимой проверки [46]. Для картирования таких целевых последовательностей на референсную последовательность генома человека была использована программа BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat) с порогом отсечения, установленным на уровне сходства 90%. Затем использовалась аннотация геномного браузера UCSC Genome Browser для генов RefSeq, мРНК и сплайсированных вариантов EST ST на референсные последовательности хромосом генома человека по сборке NCBI Build 36 (hg18). Аннотация наборов проб выполнялась по исходным целевым (таргетным) последовательностям содержащим весь набор проб, а не по отдельным 25-мерным нуклеотидных пробам. Для картирования таргетных последовательностей использовалась программа BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat), с параметром сходства минимум 90% на референсный геном человека выпуска NCBI Build 35 и 36.1 (что соответствует идентификаторам hg17 и hg18 геномных релизов).

Набор проб Affymetrix рассматривался как проблемный, если его таргетная (целевая) последовательность (1) не могла быть выровнена по BLAT ни с одним районом в геноме человека даже с 90% сходством; (2) выравнивание по BLAT показывало совпадение в различных локусах генома человека; (3) выравнивание последовательности находилось в противоположной ориентации к последовательности целевого гена (т.е. наблюдалось точное совпадение всех последовательностей экзонов и блоков таргетной последовательности набора Affymetrix, но в противоположной цепи ДНК). В сложных случаях расположения транскриптов в геноме последовательно проверялось совпадение таргетной последовательности набора проб к аннотации нуклеотидной последовательности генов RefSeq, затем к аннотации mRNA (расположение сплайсированных мРНК), затем к аннотации сплайсированных EST

(транскрипты, аннотированные с помощью технологии EST), представленной в таблицах геномного браузера UCSC. Отметим, что могут использоваться и другие базы данных аннотаций расположения генов человека в геноме, например «KnownGenes», «UCSC genes». Аннотация RefSeq использовалась как самая надежная, «ручная» аннотация известных генов, аннотация mRNA - как описание возможных изоформ генов, и аннотация EST как описание экспериментально найденной транскрипционной активности с известным расположением сплайс-вариантов.

Данные по наличию геномных повторов были получены с помощью разметки RepBase в геномном браузере UCSC (http://genome.ucsc.edu/cgi-bin/hgTracks, табл. RepeatMasker). Была построена таблица перекрывания в геноме расположения таргетной последовательности набора проб с повторяющимися геномными последовательностями, аннотированными в RepeatMasker с классификацией по семействам и типам повторов - DNA, LTR, LINE, SINE, простые повторы и повторы низкой сложности (“simple” и “low complexity”). Для каждой таргетной последовательности Affymetrix было построено описание типа геномного повтора, длина перекрывания в нуклеотидах и процент от общей длины.

Если таргетная последовательность набора проб микрочипа, предназначенная для детекции транскрипции генов человека полностью отсутствует в геноме человека (не определяется с помощью BLAT), или встречается многократно в различных геномных локусах, то такой набор проб является источником неопределенности и дает шумовой сигнал на микрочипе. Эти наборы проб должны быть исключены из анализа результатов микрочиповых экспериментов.

С помощью BLAT на геном человека (hg18) были картированы все 44,692 таргетные последовательности чипов U133A и U133B (исключая контрольные наборы искусственных олигонуклеотидных проб).

Было установлено, что 1212 (или 2.7%) таргетных последовательностей не соответствуют ни одному локусу в геноме человека (обозначено как Tag0, то есть 0 совпадений, см. Таблицу). Подавляющее большинство - 42708 (или 95.5%) таргетных последовательностей имело однозначное расположение в геноме (обозначим их Tag1, где 1 соответствует одному положению в геноме). Относительно небольшое число 772 (или 1.7%) таргетных последовательностей имело множество локализаций в геноме человека (Tag2+, где «2+» обозначает две и более локализации).

Группа наборов проб Tag2+ определяется как сумма групп Tag2, Tag3, Tag4 и т.д., где цифры 2, 3, 4 и далее обозначают число локализаций найденных программой BLAT. Группы Tag0 и Tag2+ по своей природе могут вызывать шум или сигнал кросс-

гибридизации на микрочипе. Наборы проб Tag0 относились в основном к аннотации mRNA и EST, но не были сверены с геномной ДНК, будучи ассоциированы с плохо аннотированными транскриптами, последовательностями ошибочно помеченными как "human" в базе данных GenBank во время первого дизайна проб микрочипа Affymetrix (2003 год). Например, около 45% последовательностей Tag0 были классифицированы как чужеродные последовательности ("xeno-sequence/non-human"), включая последовательности мыши, крысы, коровы, различных патогенов человека. Так, набор проб 224340_at это ген мыши c-myc с дополнительной инсерцией TGA; 217283_at картируется на Shox2 (гомеобокс) мыши; 217255_at это 100% ген коровы SQSTM1. Другие последовательности этой группы можно характеризовать как последовательности генома человека с низким уровнем сходства. Небольшое число последовательностей попало в плохо аннотированные районы (например, 222196_at располагался в не ассемблированном (*random) фрагменте хромосомы).

Таблица 2.5

Статистика числа локализаций таргетных последовательностей наборов проб Affymetrix в геноме человека

Число картирований (hg18)	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6+	Tag0	Total
Число наборов проб Affymetrix IDs	42708	450	129	67	42	84	1212	44692
Процент	95.56	1.0	0.28	0.14	0.09	0.18	2.71	100

Локализация таргетной последовательности микрочипа в геноме должна соответствовать расположению гена, экспрессию которого должен детектировать набор проб. Однако расположение самого гена может меняться от одного выпуска базы данных (геномного релиза) к другому, и в некоторых случаях положение таргетной последовательности не может быть формально определено. Например, 208303_s_at располагается на различных хромосомах: X и Y, следуя картированию гена CRLF2 (cytokine receptor-like factor 2 isoform 1). Кодирующая часть этого гена не полностью определена (статус CDS обозначен как «not complete»). Другой пример - набор проб 207353_s_at, таргетная последовательность которого картируется на не ассемблированную часть хромосомы 4 (chr4_random) следуя локализации гена HMX1 (homeo box H6 family 1). Таргетная последовательность Affymetrix 221715_at не картируется на геном человека, ни на сборку выпуска hg17, ни на hg18. Отметим, что в последнем релизе hg19 эта последовательность картируется на интрон гена KAT6A, что опять-таки не соответствует ни одной форме сплайсированной мРНК и не может измерять экспрессию никаких генов.

Были найдены множественные геномные локализации некоторых исключительно избыточных таргетных последовательностей наборов проб. Так, последовательность 81737_at имеет 22 различных локализации в геноме человека; 213089_at также имеет более 11 вариантов расположения в геноме человека. Некоторые из таких вариантов присутствуют как в сборке генома выпуска hg17, так и в сборке hg18, что говорит об исходно неправильном дизайне проб микрочипа.

Исследование присутствия мобильных элементов в нуклеотидных таргетных последовательностях наборов проб Affymetrix

Были рассмотрены данные микрочипов Affymetrix GeneChip, относящиеся к клиническим экспериментам на опухолевых тканях, хирургически полученных при лечении, в применении к анализу спонтанной экспрессии мобильных элементов [46]. Отметим, что дизайн проб этого микрочипа исходно не предназначался для такого анализа, что потребовало разработки специальных статистических оценок.

Непосредственная детекция экспрессии транспозонов, передвижений и встроек мобильных элементов в хромосомы соматических клеток технически ограничена. По методическим и техническим причинам геномные повторы обычно исключаются из дизайна микрочипов [160], в частности из-за избыточности мобильных элементов в геноме и сложности подбора уникальных проб. Таким образом, их потенциальная транскрипционная активность остается недостаточно охарактеризованной, несмотря на многочисленные наблюдения присутствия транскрипции в различных тканях при заболеваниях человека [161, 162]. С помощью технологий анализа полноразмерных «кэпированых» транскриптов была показана связь инициации транскрипции с присутствием в 5′-области генов ретротранспозонов (от 6 до 30 %) в геномах мыши и человека [392]. Экспрессия генов, содержащих ретротранспозоны в 3′-НТР, уменьшена по сравнению с генами, не содержащими таких транскриптов [392].

Мы оценили присутствие мобильных элементов генома человека в целевых последовательностях транскриптов, представленных наборами проб на микрочипе Affymetrix, для поиска ассоциации с измерением экспрессии таких проб и возможной транскрипцией классов мобильных элементов в соматических клетках. Ранее была выполнена независимая аннотация наборов проб микрочипов Affymetrix U133 на основе картирования нуклеотидных последовательностей проб на референсные последовательности генома человека [151], выявлен ряд несоответствий в аннотации наборов проб и показано, что изменения в идентификации генов могут затрагивать до 30–50 % наборов проб [157, 159]. В то же время вопрос картирования проб на целевые последовательности генов, содержащие мобильные элементы в геноме человека, не

рассматривался детально. Статистически связь присутствия последовательностей мобильных элементов и систематических изменений в экспрессии генов была показана только в работе Orlov с соавт. [46].

Примеры расположения целевой последовательности наборов проб в геноме пересекающиеся по расположению с SINE и LTR, соответственно, приведены на рисунке 2.18.

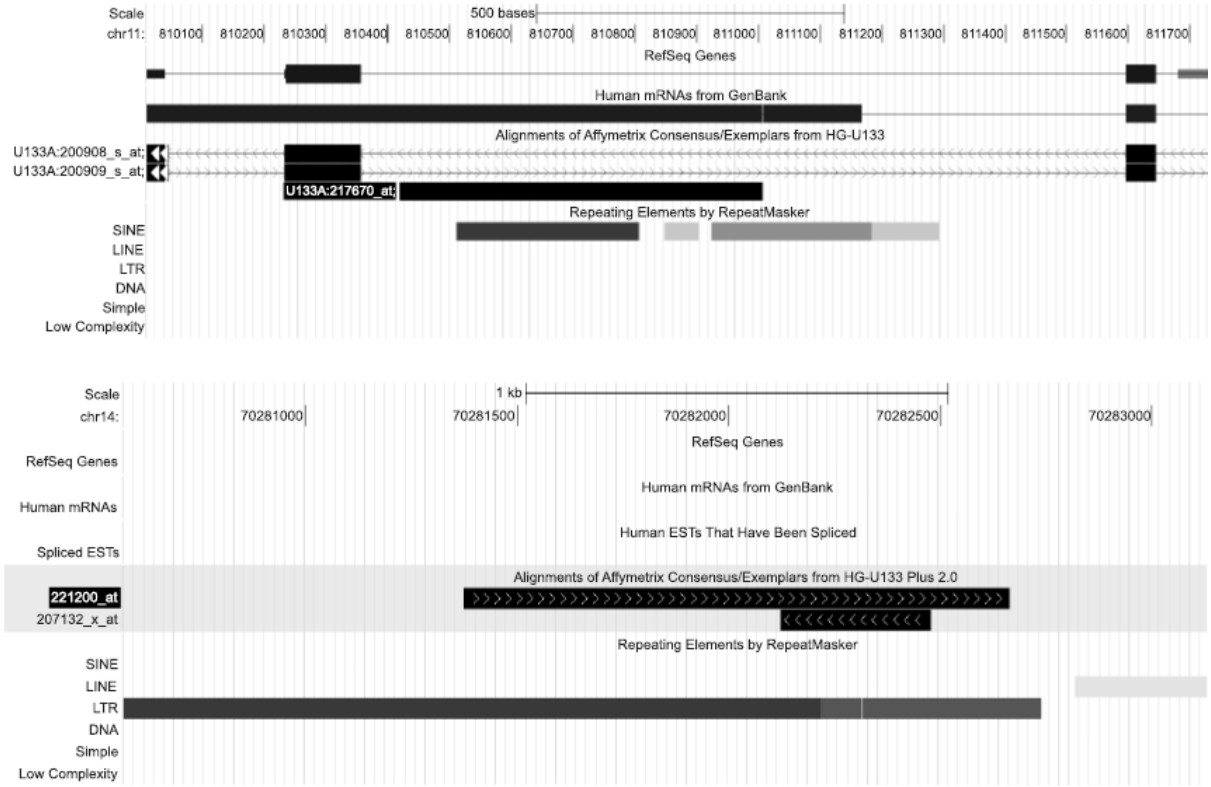


Рис. 2.18. Примеры перекрывания целевых последовательностей наборов проб с мобильными элементами в геноме человека (визуализация UCC Genome Browser).

(Верхняя панель) – пример целевой последовательности Affymetrix 217670_at на хромосоме 11 человека, совпадающей с повторяющимися элементами из (*SINE*, семейство *Alu*-повторов). Данная целевая последовательность находится в интроне. Возможна «экзонизация» соответствующего повтора в неаннотированных изоформах транскрипта (присутствие сплайсированных мРНК) на данном участке.

(Верхняя панель) – целевая последовательность набора проб 221200_at не соответствует ни генам, ни мРНК и находится внутри протяженного геномного повтора int (EVK, LTR).

Мы использовали аннотацию наборов проб Affymetrix U133 GeneChip, выполненную в работах [46, 47, 49] с целью детального изучения влияния мобильных элементов на изменение экспрессии генов в раковых тканях. Анализ был сделан на экспрессионных данных GeneChip из больших выборок (базы данных и репозитарии GEO (GeneExpressionOmnibus), http://www.ncbi.nlm.nih.gov/geo, и ArrayExpress, http://www .ebi.ac.uk/arrayexpress/) по экспрессии генов в раковых клетках, отличающихся по клиническим и генетическим параметрам, а также по степени агрессивности роста опухоли.

Проверка качества целевых последовательностей Affymetrix выполнялась последовательно: сначала были отфильтрованы некартируемые и неоднозначно картируемые последовательности, затем последовательности в неверной ориентации к аннотированным генам. Затем проводилась разметка мобильных элементов из RepBase. Для каждой целевой последовательности, однозначно картированной на геноме, была получена таблица геномных повторов, классифицированных по семействам повторов и типам (DNA, LTR, LINE, SINE включая MIR и Alu, а также простые тандемные повторы и участки низкой сложности), определены длина и процент длины, занятый геномными повторами данных типов. Суммарная статистика приведена в таблице 2.6. Как отмечалось ранее, процент «экзонизированных» геномных повторов невелик [152]. Перекрывание целевой последовательности транскрипта гена с геномным повтором не является ошибкой и не указывает на факт экзонизации.

Таблица 2.6

Классификация геномных повторов в целевых последовательностях наборов проб Affymetrix U133

Группа геномных повторов	Классы повторяющихся элементов по RepBase	Число наборов проб	Доля наборов проб, %
Короткие транспозоны (<300 п.н.)	*SINE/Alu*, *SINE/MIR*	3200	31,8
– в том числе *Alu*	*Alu*	1807	18,0
Длинные транспозоны (>300 п.н.)	*LINE/CR1*, *LINE/L1*, *L2*	2191	21,8
– в том числе *L1*	*L1*	1394	13,9
LTR	LTR/ERV1/ERVK/ERVL/MaLR	1235	12,3
DNA	*MER1*, *MER2*	1005	10,0
Другие повторяющиеся элементы и сателлитные повторы	RNA, rRNA, Satellite, scRNA, snRNA, srpRNA	52	0,5
Участки низкой сложности, простые повторы	Low_complexity	2373	23,6

Данные для анализа экспрессии

Было проанализировано распределение значений экспрессии транскриптов, полученных с помощью микрочипов Affymetrix U133A и U133B в 249 образцах первичных опухолей молочной железы (NCBI Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geo/; данные GSE4922). Выборки раковых тканей были разделены на группы, соответствующие гистологическим классам опухоли по степени агрессивности (метастазирования) рака молочной железы. Объем выборок составлял от 40 до 100 образцов [509]. Были использованы также данные экспрессии из нескольких выборок нормальных и раковых тканей мозга (GEO GDS1962), 29 наборов микрочиповых данных Affymetrix, представляющих рак легких (GEO ID: GSE5816;

http: //www.ncbi.nlm.nih.gov/geo/) [149]. Все данные прошли контроль качества сигнала экспрессии, нормализацию по алгоритму MAS5 [154]. Затем значения экспрессии были логарифмически нормированы для сопоставления экспрессии генов на микрочипах из разных экспериментов. Измерение уровня экспрессии проводилось без выделения отдельных проб.

Статистический анализ

Была использована эмпирическая функция распределения сигнала экспрессии на выборке проб микрочипов. Эмпирическая функция распределения была построена также для групп наборов проб, классифицированных по степени присутствия повторов в их целевых последовательностях. Для сравнения распределений использовалась компьютерная симуляция – были сгенерированы случайные группы (выборки) наборов проб того же размера, процедура повторялась с помощью датчика случайных чисел.

Для контроля предсказательной (диагностической) способности наборов проб, содержащих мобильные элементы и не содержащих их, были использованы опубликованные ранее результаты группировки тканей (различных гистологических классов рака молочной железы) [509]. Данные группы разделяются по экспрессии около 4000 дифференциально экспрессирующихся генов [46].

Для численной оценки использовалась программа SAM (Statistical Analysis of Microarrays) [153]. Для каждого набора проб (измерения транскрипции) данная программа рассчитывает «значимость различия» между двумя выборками данных (группами опухолей) с помощью оценки значения доли ложного предсказания FDR («false discovery rate») (параметр q-value). При фиксированном пороговом уровне q-value программа SAM идентифицирует набор генов (наборов проб), позволяющих достоверно разделить выборки и определяет дифференциально экспрессирующиеся гены. Зафиксировав значение этого параметра на уровнях 0,05 и 0,015, мы оценили фракцию наборов проб, позволяющих дискриминировать типы опухолей и содержащих при этом мобильные элементы. Дискриминирующая способность проб, содержащих мобильные элементы, была оценена с помощью отношения наблюдаемой и ожидаемой доли проб, содержащих мобильные элементы. Для оценки значимости результатов использовался критерий Манна-Уитни (U-test) и односторонний точный критерий Фишера для таблиц данных.

Статистика геномных повторов в наборах проб

В целом до 25 % целевых последовательностей для наборов проб проявляют значимое сходство с мобильными элементами (геномными повторами), распространенными в геноме человека [46].

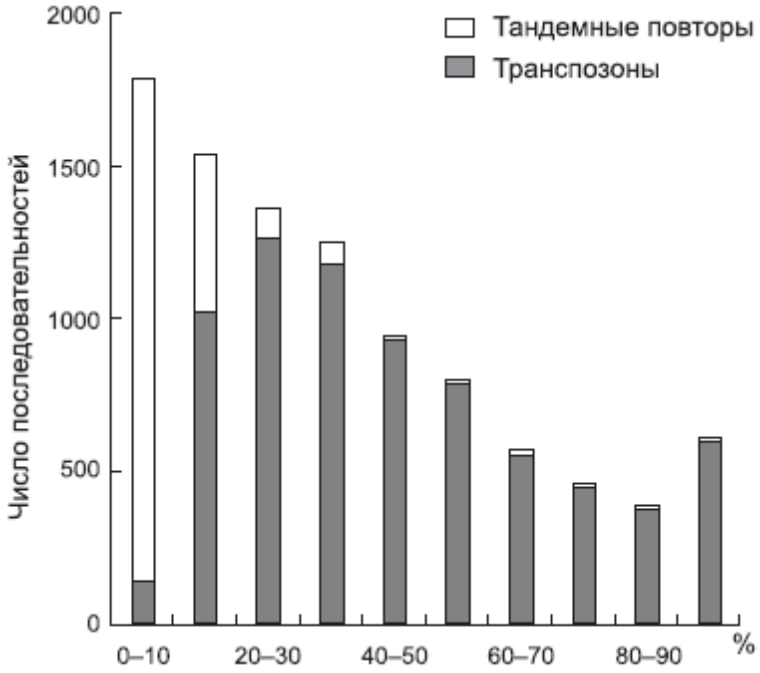


Рис. 2.19. Распределение числа целевых (таргетных) последовательностей наборов проб Affymetrix U133, пересекающихся с повторяющимися элементами в геноме человека в зависимости от процентной доли геномных повторов в последовательности (ось абсцисс – от 0 до 100 %) [49].

Большое число целевых последовательностей наборов проб на микрочипе проявляют значимое сходство с геномными повторами, представляя, тем не менее, лишь малую часть от более чем 5 млн. участков, размеченных RepeatMasker [77] в геноме человека (таблица RepeatMasker геномного броузера UCSC). Напомним, что размеры последовательностей варьируют от 100 до 500 п.н. Доля перекрывания, как правило, невысока, менее 50% для большинства последовательностей. В то же время, несколько тысяч целевых последовательностей перекрываются с геномными повторами более чем на 40 % от своей длины, а около 600 целевых последовательной – более чем на 90 %, что, несомненно, влияет на качество сигнала. Такое перекрывание таргетных последовательностей позволяет детектировать экспрессию мобильных элементов, а не генов, для которых исходно предназначался дизайн наборов проб. При этом часть последовательностей содержит простые тандемные повторы и участки низкой сложности, занимающие менее 10 % от общей длины целевой последовательности, что не должно оказывать влияние на сигнал экспрессии (рис. 2.18).

Классификация целевых последовательностей Affymetrix по качеству дизайна и соответствию аннотации генов

Таблица 2.7 содержит общую статистику различных категорий неверно определенных целевых последовательностей проб Affymetrix U133 на основе геномной сборки hg18. Около 6 % составляют последовательности, картируемые на геном в

различных участках (неоднозначно картируемые), не соответствующие последовательностям генома человека, и последовательности, картируемые в противоположной ориентации к транскрибируемым последовательностям генов. Общая фракция целевых последовательностей наборов проб велика, до 25 %, но большей частью геномные повторы «присутствуют» в таких последовательностях лишь частично, и их можно считать адекватными для измерения экспрессионного сигнала. В целом 86 % наборов проб были рекомендованы к использованию (Orlov et al., 2007).

Таблица 2.7

Общая классификация проблемных целевых последовательностей микрочипа Affymetrix U133

Группа целевых последовательностей наборов проб	Число	Доля, %
Неоднозначно картируемые на геном человека	1984	4,4
Tag0	1212	2.71
Tag2+	772	1.72
Картированные в обратной ориентации к транскрипту	810	1,8
Перекрывающиеся с геномными повторами, в том числе:	3387	7,6
80–100 % длины последовательности	761	1,7
60–80 %	936	2,1
40–60 %	1690	3,8
Итого не рекомендуется использовать	**6181**	**13,8**
Общее число корректных наборов проб (включая перекрывание менее 40 % длины последовательности), в том числе:	**38511**	**86,2**
Частичное перекрывание с транскриптами в противоположной цепи	13260	29.66
Неправильная ориентация по отношению к EST	487	1.08
Таргетная последовательность набора проб с 20-40% геномных повторов	2409	5.39
Таргетная последовательность с менее чем 20% повторов	1210	2.7
Общее число последовательностей микрочипов U133A и B	44692	100

Из таблицы видно, что критериям качества таргетных последовательностей в геноме человека удовлетворяют только 86,2% наборов проб. При этом возможно частичное перекрывание с транскриптами в противоположной цепи ДНК, что не является ошибкой дизайна таргетных последовательностей, а отражает сложную структуру самих транскриптов. Только 1.8% таргетных последовательностей картированы в геноме полностью в обратной ориентации, что является ошибкой их дизайна при исходной разработке микрочипа, и поэтому не должны быть использованы

Исходно дизайн микрочипа не предназначался для исследования экспрессии мобильных элементов. Но эти данные могут быть использованы для статистических оценок, так же, как и для фильтрации и калибровки измерения экспрессионного

сигнала. Эти данные и таблицы оценок для всех наборов проб представлены в компьютерном ресурсе АРМА для оценки качества проб на микрочипе платформы Affymetrix U133 [46, 47].

На следующей таблице представлено сравнение чипов A и B той же платформы Affymetrix. Чип B был спроектирован позднее и включает меньшее число белок-кодирующих генов, а большее - транскриптов мРНК. Поэтому имеет смысл оценивать качество дизайна таргетных последовательностей и статистику экспрессии генов на этих чипах по отдельности.

Таблица 2.8

Сравнение качества геномной аннотации наборов проб чипов U133A и U133B платформы Affymetrix

Наборы проб	Число наборов проб	Число корректных наборов проб, прошедших контроль качества	Процент корректных наборов проб, прошедших контроль качества
Пробы из пересечения A и B	100	98	98.0
Служебные наборы проб	68	-	-
Чип U133A	22115	19753	89.3
Чип U133B	22477	18660	83.0

Таблица показывает число и процент корректных таргетных последовательностей чипов U133A и U133B. Видно, что доля таргетных последовательностей наборов проб, прошедших контроль качества (т.е. уникальность локализации в геноме, корректная ориентация относительно генов, невысокий процент перекрывания с геномными повторами) выше для чипа U133A (около 89%) против 83% для U133B.

Сравнение средних значений экспрессии наборов проб, содержащих геномные повторы

Выполнено сравнение средних уровней экспрессии для групп целевых последовательностей, содержащих геномные повторы в зависимости от длины последовательности (с шагом гистограммы 10 %), и для корректных целевых последовательностей (норма). Для каждой группы наборов проб, целевые последовательности которых содержат повторы, мы определили средние значения сигнала гибридизации наборов проб (в логарифмической шкале) и коэффициент вариации (дисперсия, нормированная на среднее значение) на выборках данных опухолей. На рис. 2.20 показано уменьшение среднего значения сигнала при увеличении доли геномных повторов в целевой последовательности. В то же время

коэффициент вариации имеет противоположный тренд и может быть достаточно большим, более 0,1, для последовательностей, почти полностью занятых геномными повторами.

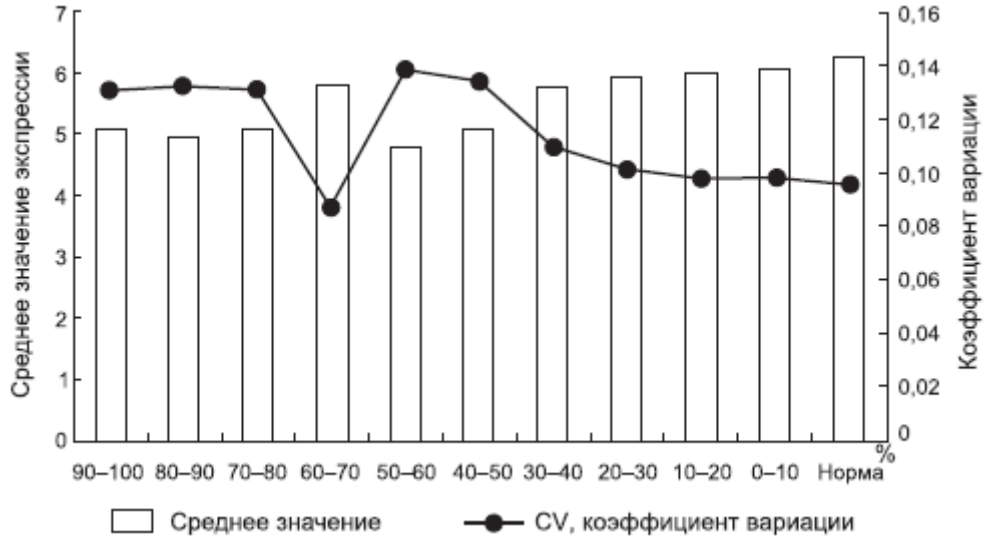


Рис. 2.20. Среднее значение экспрессии и коэффициент вариации наборов проб, перекрывающихся с геномными повторами, по выборке опухолевых тканей.

Ось абсцисс – группа целевых последовательностей, занятых повторами на 90–100 %, 80–90 % и т. д. вплоть до 0 %, корректно определенных последовательностей (норма). По оси ординат слева – среднее значение экспрессии соответствующей группы в логарифмической шкале сигнала гибридизации (колонки гистограммы), справа – коэффициент вариации (линия), безразмерное значение. Видны противоположные тренды – уменьшение среднего значения экспрессии при увеличении доли геномных повторов и увеличение коэффициента вариации (зашумленности сигнала). Данные приведены по выборке образцов опухолей молочной железы (гистологический Grade I).

Способность целевых последовательностей, перекрывающихся с геномными повторами, к определению дифференциально экспрессирующихся генов

Было выполнено сравнение способности наборов проб дискриминировать дифференциально экспрессирующиеся гены в выборках образцов опухолей различных типов. Гистологически опухоли молочной железы классов II и III (низко- и высокометастазирующие) различаются, что может быть статистически на микрочипах показано дифференцированной экспрессией нескольких тысяч наборов проб Affymetrix. Используя программное обеспечение SAM [153], мы отобрали набор из 6144 дифференциально экспрессирующихся генов на микрочипах U133A&B на фиксированном уровне q-value ложного положительного предсказания, не превышающем 1,5 % .

Предполагая, что перекрывание с мобильными элементами приводит к ухудшению качества сигнала из-за неспецифического связывания проб с посторонними

транскриптами и не может быть использовано для дискриминации, мы вправе ожидать, что такие наборы проб должны быть недопредставлены во множестве дифференциально экспрессирующихся генов. Было рассчитано число наборов проб, занятых повторами на 10, 20, … , 100 %, найденных в данном множестве дифференциально экспрессирующихся.

Доля дифференциально экспрессирующихся наборов проб, для которых соответствующая целевая последовательность перекрывается с геномными повторами в геноме человека, может быть рассчитана по формуле:

$$r = (Rs / R)/(Ns / N),$$

где N – общее число наборов проб микрочипа; R – число наборов проб, перекрывающихся с геномными повторами, Ns – число наборов проб, дифференциально экспрессирующихся в исследованных видах опухолей по статистическому тесту программы SAM; Rs – число наборов проб, перекрывающихся с геномными повторами и дифференциально экспрессирующихся в тех же видах опухолей по тесту SAM.

Таким образом, r – это отношение наблюдаемой доли наборов проб, связанных с геномными повторами, к ожидаемой доле, ее значение может быть как больше, так и меньше единицы. Было определено, что присутствие простых тандемных повторов и участков низкой сложности не влияет на способность наборов проб дискриминировать опухоли. Простые повторы в целом занимают незначительную часть целевых последовательностей (см. рис.), наборы проб, ассоциированные с простыми повторами, имеют малую вариабельность сигнала (коэффициент вариации по клиническим выборкам данных) и функционально не влияют на дискриминирующие свойства наборов проб при сравнении биологически различных выборок образцов тканей.

В то же время наборы проб, целевые последовательности которых содержат последовательности мобильных элементов, существенно хуже дискриминируют опухоли ($r < 1$). Более того, наблюдается тренд изменения дискриминирующего параметра r в зависимости от доли целевой последовательности, занятой мобильными элементами, особенно для протяженных геномных повторов (LTR и LINE): чем больше геномных повторов присутствует в целевой последовательности набора проб микрочипа, тем меньше отношение r (рис. 2.21).

Статистический анализ данных экспрессии наборов проб, связанных с геномными повторами, на выборках образцов опухолей (молочной железы, тканей мозга) выявил общую воспроизводимую тенденцию: 1) увеличение шума в сигнале экспрессии (коэффициента вариации); 2) уменьшение среднего уровня сигнала

экспрессии и 3) увеличение числа ложных корреляций, не связанных с взаимной регуляцией транскрипции. Таким образом, при интерпретации данных экспрессионных микрочипов необходим учет особенностей геномной аннотации мобильных элементов.

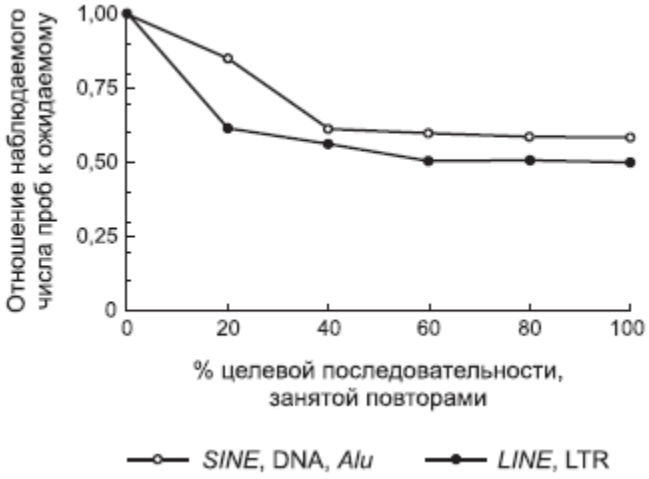


Рис. 2.21. Оценка отношения r дискриминирующих наборов проб к ожидаемому (при фиксированном q < 1,5 %) как функция процента целевой последовательности, занятой транспозонами [49].

Оценка активности мобильных элементов в геноме человека может быть выполнена на основе различных технологий, включая полное ресеквенирование индивидуальных геномов и секвенирование индивидуальных транскриптом [6]. В перспективе это позволит более полно описать картину активации мобильных элементов в соматических клетках при повреждающих воздействиях различных типов, описать функционально активные копии и места встраивания мобильных элементов в геноме. Отметим, что новые технологии имеют ряд технических проблем, в частности достаточно высокий уровень ошибок секвенирования, чувствительность к GC-составу и гетерогенности последовательностей [14]. Это требует разработки специализированных компьютерных методов анализа.

Заключение к разделу

Источники шума в численных данных микрочиповых экспериментов могут быть различны, связаны как с технологическими причинами, так и с неверной интерпретацией (аннотацией) наборов проб [148, 151, 156]. Критика качества наборов проб, связанных с присутствием в них участков сходства с геномными повторами, была высказана ранее, однако без статистических оценок применительно к экспрессии мобильных элементов. Здесь показаны статистические оценки влияния повторов на интенсивность экспрессионного сигнала, коэффициент вариации и способность дискриминировать различные биологические классы (число дифференциально экспрессирующихся наборов проб) на больших выборках клинических данных [46].

Несмотря на ошибки дизайна проб и аннотации последовательностей-мишеней Affymetrix U133A&B, технологически платформа дает воспроизводимые результаты, подтвержденные большим объемом данных.

В целом, потенциал микрочиповых данных может быть использован гораздо полнее через интеграцию геномной аннотации и клинических данных. Дальнейшее уточнение аннотации референсного генома, в частности, в связи с новыми проектами ресеквенирования генома человека, может увеличить число неверно аннотированных проб и привести к переоценке данных, накопленных при использовании данного типа микрочипов за последние годы [159].

Детекция экспрессии мобильных элементов на данном типе микрочипа не была спланирована первоначально и показана только как результат статистического анализа. Исследование транскрипции в геноме человека с помощью новых технологий секвенирования, в частности RNA-seq, позволяет найти новые транскрипты в геноме, детекция которых невозможна с помощью микрочипов [392]. Таким образом, измерение уровней экспрессии мобильных элементов в соматических клетках, в частности в опухолевых тканях, может быть сделано с помощью других подходов, но, к сожалению, теряется огромный накопленный массив клинических данных микрочипов.

Отметим, что гибридизационный сигнал от набора проб с обнаруженным перекрыванием целевой последовательности с каким-либо геномным повтором из RepBase, размеченным с помощью RepeatMasker, не дает информации о транскрипции конкретно этого мобильного элемента или другого гомологичного ему элемента, расположенного на других хромосомах. Таким образом, мы можем сравнивать только классы мобильных элементов и оценивать статистически их влияние на сигнал экспрессии.

Как показано в Главе 1, экспонирование клеток к ДНК-повреждающим воздействиям, таким, как лекарства химиотерапии или радиация, может вести к индукции транскрипции SINE-элементов, что подтверждает глобальную активацию транспозонов в геноме при стрессовых условиях [393]. Есть данные об экспрессии в соматических клетках элементов семейства L1 (LINE) [388-391]. Механизмы воздействия экспрессии мобильных элементов могут не ограничиваться встройками ДНК и структурными изменениями генома. Показано, что РНК, транскрибируемая с Alu-повторов, может взаимодействовать с РНК полимеразой II и подавлять экспрессию некоторых белок-кодирующих генов [394]. Таким образом, активация транспозонов в геноме может вести к изменению экспрессии генов в раковых тканях и при повреждающих воздействиях, что требует дальнейшего изучения.

2.6. База данных RatDNA специализированных микрочипов генов крысы

Анализ экспрессии генов может выполняться и на специализированных микрочипах, содержащих до нескольких сот генов, значительно меньше, чем микрочипы высокой плотности, такие как Affymetrix. Такой микрочип был разработан в ИЦиГ СО РАН для исследовании экспрессии группы генов на модельных животных – крысах [53].

Фенотипическим проявлениям катаракты и возрастной макулярной дегенерации (ВМД) предшествуют изменения экспрессии генов, однако вклад изменений транскриптома в процесс нормального физиологичного старения и тем более в развитие этих заболеваний, особенно на ранних стадиях, остаётся не ясным в силу сложности проведения таких исследований на людях. Исследования транскриптома проводятся и на моделях ВМД – на животных, развитие ретинопатии у которых вызывают, как правило, воздействием различных физических факторов (УФ - или лазерным излучением, гипероксией), которое только частично воспроизводит картину развития ВМД [510]. Систематическое исследование раннего развития заболевания невозможно на доклинических стадиях, на пациентах, что обуславливает необходимость испытаний на лабораторных животных. Соответственно необходима систематизация такой экспериментальной информации в базах данных.

Для систематизации информации об экспрессии генов, связанных с ВМД, в рамках работ по Технологической платформе «Медицина будущего» в ИЦиГ СО РАН был разработан ДНК-чип для исследований экспрессии генов крыс, создана база данных экспрессии генов и веб-портал с ассоциированной информацией по данной проблеме [53].

Стартовой страницей портала является страница с информацией по проекту исследования заболеваний старения на лабораторных животных (http://pixie.bionet.nsc.ru/ratdna/index.php). С главной страницы можно совершить переход в соответствующие разделы: «Общая информация о проекте», «Этапы», «Результаты», «Литературные источники», «Рабочий сайт проекта».

Портал разработан на языке PHP, база данных RatDNA разработана на MySQL. Сервер MySQL управляет доступом к данным, позволяя работать с ними одновременно нескольким пользователям, обеспечивает быстрый доступ к данным.

База данных генов крысы для микрочипа

При переходе в раздел базы данных RatDNA (рис. 2.22) в навигационном меню доступен раздел, в котором представлена справочная информация по таблице RatDNA-chip. Так же по таблице можно осуществить поиск, введя название искомого гена в

поле поиска. Помимо возможности просмотра и работы с таблицами на сайте, они доступны для загрузки.

Компьютерный портал анализа экспрессии генов крысы

Главная | RatDNA | Help | RatDNA-OMIM | RatDNA-AMD | ЦКП "Биоинформатика" | Функциональное описание селектированных генов

Набор генов крысы, связанных с заболеваниями старения

<Скачать таблицу>

№	Идентификатор RGD_ID	Символ гена	Описание гена	Хромосома	Начало	Конец	ID транскрипта	Ориентация гена в геноме	Число экзонов	
1	68358	Acan	aggrecan	1	134787341	134848992	NM_022190	+	18	CCAACACCTACAAGCACA
2	1305051	Aen	apoptosis enhancing nuclease	1	134615998	134625367	NM_001108487	+	4	agtgtactgtgagaaatcagcttgtttgtgc
3	619885	Ak3	adenylate kinase 3	1	232658879	232684083	NM_013218	-	5	TTTCCTAAGACTTCTCTGA
4	620844	Apba1	amyloid beta (A4) precursor protein-binding, family A, member 1	1	227106828	227309416	NM_031779	+	13	ATAACCACTGGCAGGTAC
5	620845	Apba2	amyloid beta (A4) precursor protein-binding, family A, member 2	1	118970882	119156605	NM_031780	+	14	ATGTATAATGATGACCTTA
6	2122	Apbb1	amyloid beta (A4) precursor protein-binding, family B, member 1 (Fe65)	1	163282918	163299333	NM_080478	-	13	tgtttgaggtggagcaggaggaactggtc
7	628763	Aqp11	aquaporin 11	1	154973796	154983962	NM_173105	-	3	TTGTTCTTTTGAGTGATGT
8	2154	Arnt2	aryl hydrocarbon receptor nuclear translocator 2	1	140535823	140646838	NM_012781	-	19	TGAATGTTCTGTATGACTA

Рисунок 2.22. Фрагмент интерфейса. Таблица генов крысы и олигонуклеотидных проб «RatDNA-chip» [53].

Методы поиска генов включали процессинг данных экспрессии генов на микрочипах в тканях сетчатки глаз и ретины, опубликованные в литературе. Анализ баз данных и литературных источников и QTL-анализ [45, 53] позволил установить несколько списков генов, дифференциально экспрессирующихся в сетчатке глаза, которые были использованы при дизайне специализированного ДНК чипа. Было отобрано 113 генов, подобраны олигонуклеотидные зонды. База данных RatDNA содержит информацию как об этом наборе генов микрочипа.

На странице портала представлена таблица RatDNA-AMD. В ней собрана информация по генам крысы связанным с возрастной макулярной дегенерацией (ВМД, или AMD в англ. аббревиатуре). По таблице также можно осуществить поиск, введя название искомого гена в поле поиска, и она доступна для скачивания. Данные по экспрессии генов в ткани ретины крыс, полученные с помощью специализированного микрочипа, разработанного в ИЦиГ СО РАН, находятся на странице «RatDNA-Экспрессия генов».Объектами в базе данных являются гены крысы, их нуклеотидные последовательности и функциональная аннотация.

База данных в целом включает 5 таблиц:

(1) Таблица генов крысы RatDNA-chip предназначена для описания генов и олигонуклеотидных проб для микрочипа.

(2) Таблица генов крысы и гомологичных генам человека, связанных с наследственными заболеваниями человека «RatDNA-OMIM» предназначена для исследования ассоциаций заболеваний старения на крысах с аналогичными заболеваниями человека

(3) Таблица генов крысы и соответствующих генов человека, ассоциированных с возрастной макулярной дегенерацией «RatDNA-AMD» построена на основе анализа литературных данных и предназначена для последующего изучения экспрессии генов на ДНК-чипах и полногеномных данных транскриптомного секвенирования.

(4) Таблица «Группа генов» содержит списки генов, селектированных по дифференциальной экспрессии в тканях крысы построена в результате анализа экспериментальных данных микрочипов и является производной для анализа генных онтологий.

(5) Таблица «Экспрессия» содержит экспериментальные данные, полученные с помощью специализированного ДНК-чипа по генам крысы из таблицы RatDNA-chip.

Связи между таблицами БД RatDNA осуществляются по RGD идентификатору гена крысы (рис. 2.23).

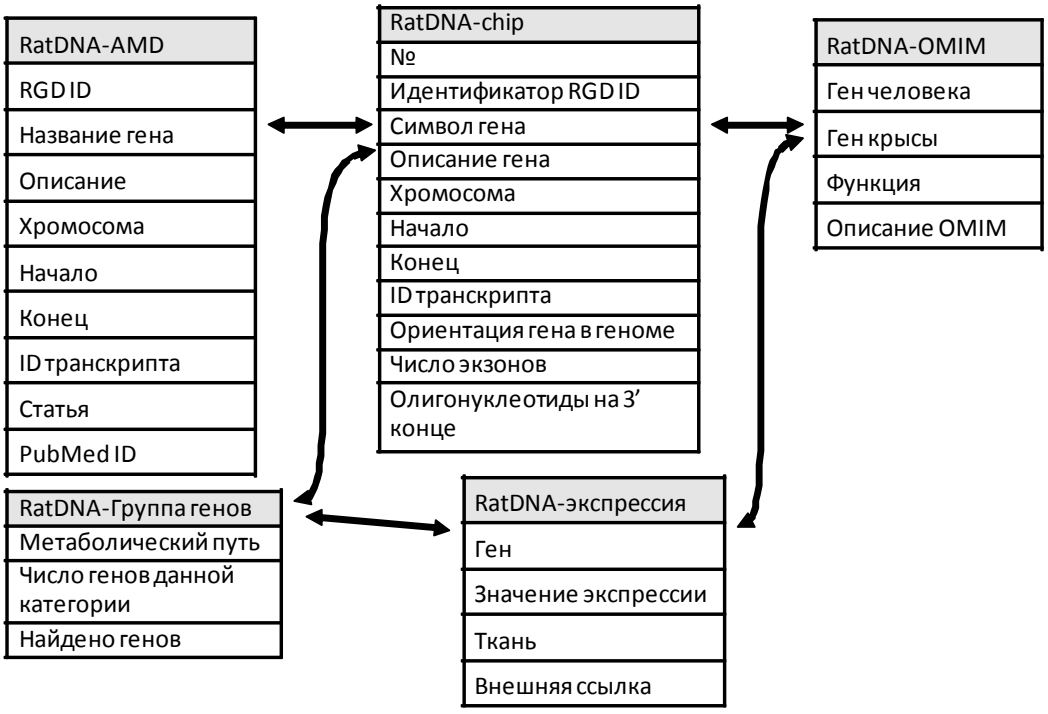


Рис. 2.23. Структура базы данных RatDNA и связь таблиц.

Исследование функций генов крысы, представленных в базе данных

Выбор генов, ассоциированных с заболеваниями старения, выполнялся по литературным данным, представленным в базах данных GEO NCBI (http://www.ncbi.nlm.nih.gov/gds) и OMIM (Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/omim). При сравнении данных OMIM по генам человека и генам крысы использовалось соответствие идентификаторов (например, HIF1A у человека и Hif1a у крысы). Использовались публикации, содержащие данные по экспрессии генов в тканях крысы, данные БД Retinobase [511].

Для анализа функций выбранных генов, относящихся к заболеваниям старения человека, было выполнено сравнение списка генов крысы с генами человека, описанными в базе данных наследственных заболеваний OMIM. Было выделено 254 категории, относящиеся к старению. По названиям генов было установлено соответствие, найдены гены, связанные с оксидативным (окислительным) стрессом, например Hif1a. Данные соответствия с OMIM представлены в таблице 2.9 («RatDNA-OMIM»).

Таблица 2.9

Соответствие найденных генов крысы и генов человека, связанных с заболеваниями старения и продолжительностью жизни (по базе данных OMIM).

Ген человека	Ген крысы	Функция	Описание OMIM
ARNTL	Arntl	Ген связан с циркадными ритмами, экспрессируется в ретине у мыши	*602550. ARYL HYDROCARBON RECEPTOR NUCLEAR TRANSLOCATOR-LIKE
BAD	Bad	Регуляция апоптоза	*603167. BCL2 ANTAGONIST OF CELL DEATH
BCL2	Bcl2	Онкоген	+151430. B-CELL CLL/LYMPHOMA 2
COQ7	Coq7	Регуляция базовых метаболических процессов, включая биосинтез, дыхание, продолжи-\тельность жизни у *C. elegans*	*601683. COQ7, S. CEREVISIAE, HOMOLOG OF
HIF1A	Hif1a	Фактор ответа на гипоксии	*603348. HYPOXIA-INDUCIBLE FACTOR 1, ALPHA SUBUNIT
IGF1R	Igf1r	Рецептор ростового фактора	*147370. INSULIN-LIKE GROWTH FACTOR I RECEPTOR
POLG	Polg	Комплекс транскрипции	*174763. POLYMERASE, DNA, GAMMA
SIRT3	Sirt3	Митохондриальная деацетилаза. Семейство белков-сиртуинов	*604481. SIRTUIN 3
TPH1	Tph1	Триптофан гидроксилаза, биосинтез серотонина	*191060. TRYPTOPHAN HYDROXYLASE 1

Был проведен анализ функций генов крысы, представленных в таблице RatDNA-chip и на микрочипе с помощью категорий генных онтологий. Для проанализированных генов было установлено соответствие 98 идентификаторов геномной аннотации RefSeq. С помощью ресурса анализа генных онтологий PANTHER (http://www.pantherdb.org) была выполнена оценка обогащенности данной группы генов категориями, относящимися к метаболическим путям, молекулярным функциям и биологическим процессам. Результаты для метаболических путей также представлены в таблице на странице БД RatDNA.

Интересно отметить присутствие категорий связанных с оксидативным стрессом (Hypoxia response), передачей сигнала FGF (FGF signaling pathway), а также метаболическими путями белков, вовлеченных в болезнь Альцгеймера и болезнь Паркинсона, развивающимися с возрастом.

Набор этих генов был протестирован также на предмет выявления регуляторных и белок-белковых взаимодействий по базе данных SPRING (Search Tool for the Retrieval of Interacting Genes/Proteins) (http://string-db.org/). Было выявлено большое число взаимодействий между белками исследуемой группы, реконструирована генная сеть. Выявлено несколько узлов сети (не менее 4-х контактов); такими узлами являются Bcl2, Bax, Timp3, Nos3, Hif1a, Igfr1, Fgdr2, Epas1, Usp3.

Интересно отметить, что многие из этих генов человека относятся к генам, связанным с заболеваниями старения по базе данных наследственности человека OMIM. Так, BCL2 (B-cell Cell/Lymphoma 2)– это известный онкоген, HIF1A (Hypoxia-Inducible Factor 1, Alpha subunit) – фактор ответа на гипоксию (недостаток кислорода), Igfr1 (Insulin-Like Growth Factor I Receptor) также является онкогеном.

Исследование экспрессии изучаемых генов на микрочипе для крыс исследуемой линии OXYS и контрольной Вистар (Wistar) показало различие уровней экспрессии в 1.3-1.4 раза (по четырем репликам), что является достаточно небольшим диапазоном (рис. 2.24). Наименьшие значения соотношения уровней экспрессии в исследуемых группах (понижение уровня экспрессии) имеют гены Stx5 (в 0.732 раза), Picalm (0.748), Prdx5 (0.75), наибольшие соотношения - гены Cd36 (1.287 раза), Nos3 (1.309) и Ucp3 (1.453). Данные этих экспериментальных измерений также представлены в базе данных RatDNA генов крысы, ассоциированных с заболеваниями старения.

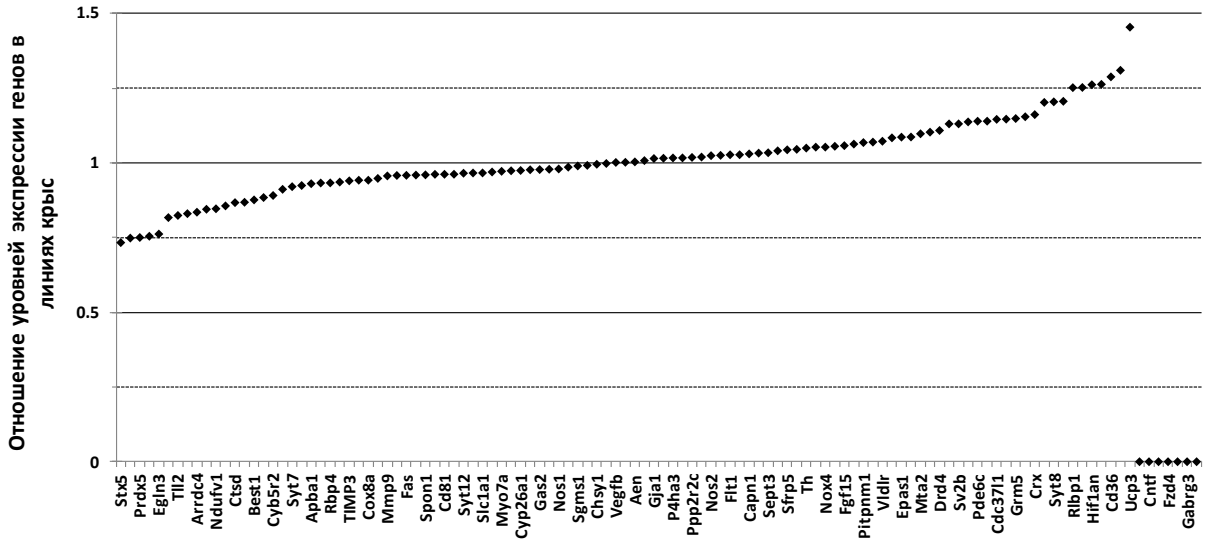


Рис. 2.24. Соотношение уровней экспрессии генов OXYS/ Wistar на разработанном чипе (ось абсцисс – гены крысы на микрочипе, ось ординат – нормализованное соотношение уровней экспрессии в тканях линий крыс) [53].

Разработанная интернет-доступная база данных интегрирует геномные данные с функциональной аннотацией генов и их связи с заболеваниями старения, включая возрастную макулярную дегенерацию, с экспериментальными данными об экспрессии этих генов в тканях лабораторных животных – крыс. Использование ссылочных таблиц на экспериментальные данные, полученные с помощью микрочипов, позволяет соотносить экспрессию генов крысы с их функцией, ролью в заболеваниях, отметить гомологичные гены в других модельных организмах. Существует возможность расширения созданной базы данных по микрочипам на исследования экспрессии генов с помощью других технологий. [45, 53].

2.7. Модели регуляторных районов транскрипции включающие антисенс транскрипты

Сложные, перекрывающиеся, структуры расположения транскрибируемых последовательностей в геноме важны для исследования глобального феномена геномной регуляции транскрипции. Автором были разработаны программы и база данных антисенс-транскриптов в геноме человека (USAP - United Sense-Antisense Pairs), использовавшаяся для анализа клинических экспрессионных данных на микрочипах [46, 48].

В целом возможно несколько вариантов перекрывающегося расположения пары цис (близлежащих) транскриптов в противоположных цепях ДНК (цис-антисенс-транскрипты) в геноме. Рисунок 2.25 показывает варианты расположения перекрывающихся транскриптов в противоположной ориентации (антисенс) в геноме.

Показано варианты взаимного расположение: перекрывание 5’-районами (стартами) - «голова-к-голове», перекрывание 3’- районами (концами) - «хвост-к-хвосту», а также вложенный вариант перекрывания, когда один транскрипт расположен внутри другого.

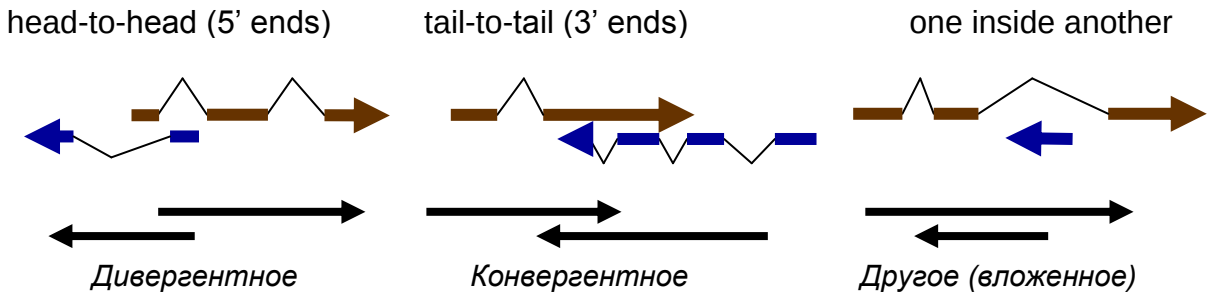


Рис. 2.25. Варианты расположения перекрывающихся транскриптов в противоположной ориентации (антисенс) в геноме. Показано взаимное расположение (слева направо):

Вариант перекрывания 5’-районами (стартами) - «голова-к-голове» (head-to-head). Дивергентное (расходящееся) расположение двух транскриптов.

Вариант перекрывания 3’- районами (концами) - «хвост-к-хвосту» (tail-to-tail). Конвергентное (сходящееся) расположение двух транскриптов.

Вариант перекрывания один внутри другого. Вложенное расположение двух транскриптов.

В действительности расположение транскриптов может иметь еще более сложный характер и включать три транскрипта в противоположных ориентациях, перекрывающихся друг с другом, как представлено на рисунке 2.26 для генов WDR6, DALRD3 и C3orf60 в геноме человека.

Рис. 2.26. Пример комплекса трех цис антисенс-транскриптов в геноме человека.

С помощью собственных программ был проведен компьютерный анализ взаимного расположения всех аннотированных транскриптов в геноме человека [48]. Использовалась аннотация генов RefSeq, аннотация транскриптов mRNA (запись в Table Browser) и аннотация EST в геноме человека.

Были собраны компиляции данных пар цис антисенс-транскриптов в парах RefSeq- RefSeq, RefSeq- mRNA, mRNA-mRNA (обозначения соответствуют источникам данных для первого и второго транскрипта в паре). Использовались также компиляция данных SATU (Sense-Antisense Transcript Units) и NATSDB [48]. Таблица 2.10 содержит варианты аннотации, рассчитанные автором, и их сравнение с опубликованными данными пар цис антисенс-транскриптов в геноме человека.

Таблица 2.10

Варианты аннотации пар цис антисенс-транскриптов в геноме человека по RefSeq, mRNA, EST

Источник данных Аннотация пары транскриптов	**SATU**	**Собственный метод**	**NAT**	**Всего без повторов**
RefSeq/RefSeq	2	1161	303	1161
mRNA-mRNA	1419	9248	282	10047
EST-EST	507	0	173	672
RefSeq/mRNA	398	7926	1119	8177
RefSeq/EST	484	0	1379	1745
mRNA/EST	1588	0	449	1980
Общее число пар*	4398	9640	3705	23782

*Примечание. Из-за постоянной корректировки аннотаций транскриптов в выпусках (релизах) баз данных NCBI, последний пересчет антисенс пар может отличаться по общему числу пар (также как и общее число генов в геноме может меняться от выпуска к выпуску аннотации генома человека и базы RefSeq).

Несмотря на распространённость цис-антисенс транскриптов в геноме человека, перекрывание транскриптов не нарушает функционирования генов, поскольку перекрывание происходит, как правило, некодирующими частями и занимает относительно небольшую долю длины гена.

Результаты расчетов длин перекрываний антисенс пар представлены в следующей таблице. Таблица 2.11 показывает, что общий размер перекрывания антисенс пар в геноме человека составляет не более 5% их размера (размер участка, занимаемый в хромосомных координатах, меньше суммы их длин).

Таблица 2.11

Размер пар антисенс транскриптов и размер пересечения транскриптов в противоположной ориентации (для аннотаций RefSeq, mRNA, EST)

Тип перекрывания пары	Аннотация транскриптов	Средний общий размах пары в геноме, Кб	Средний размер антисенс-перекрывания, Кб	% от общего размера пары
Конвергентный	RefSeq-RefSeq	75	3.4	4.6
Дивергентный	RefSeq-RefSeq	68	1.7	2.6
Вложенный	RefSeq-RefSeq	97	14.1	14.6
Конвергентный	RefSeq/mRNA	102	9.6	9.5
Дивергентный	RefSeq/mRNA	74	2.2	3.0
Вложенный	RefSeq/mRNA	109	6.7	6.2

Для дивергентного типа перекрывания, когда транскрипция идет в разные стороны с общего промоторного участка, перекрывание составляет меньший процент, до 3% от общего размера пары, чем для конвергентного типа перекрывания.

Вложенный антисенс транскрипт располагается, как правило, в интроне и занимает порядка 15% от длины большего транскрипта в паре.

Распределение размера перекрывания пар цис-антисенс транскриптов в геноме человека в нуклеотидах для конвергентного и дивергентного типа пар показано на следующем рисунке 2.27.

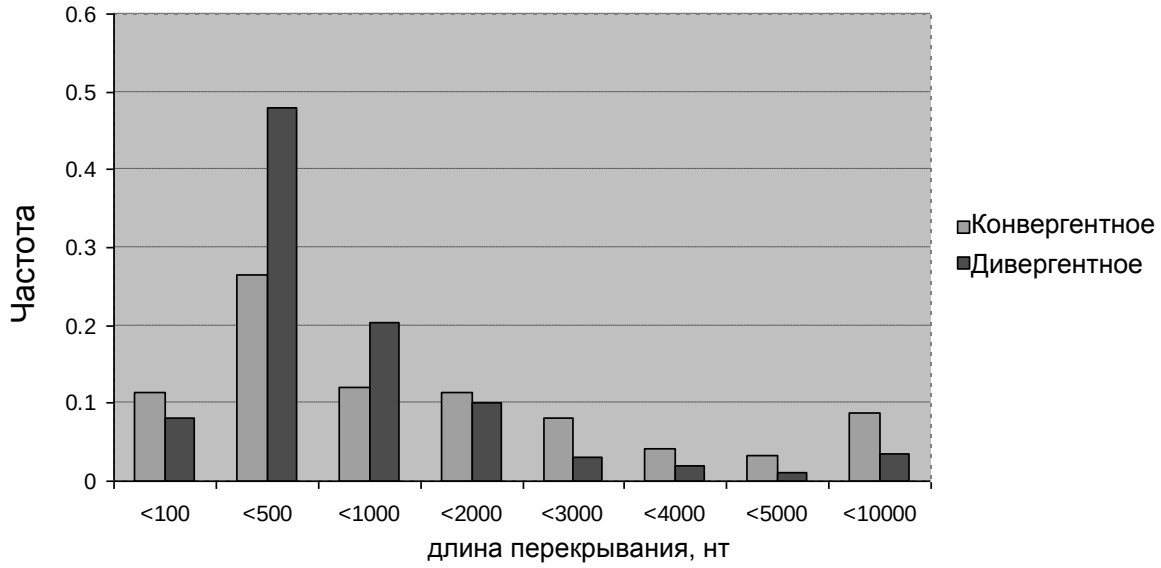


Рис. 2.27. Распределение размера перекрывания цис-антисенс транскриптов в геноме человека для конвергентного и дивергентного типа пар, в нуклеотидах.

Из рисунка видно, что размер перекрывания мал, мода распределения составляет не более 500 нт. Дивергентные пары транскриптов имеют более короткое перекрывание в геномных координатах.

Были рассчитаны ранговые корреляции экспрессии для цис-антисенс пар белок-кодирующих генов по клиническим данным микрочипов (всего около тысячи пар, см. таблицу пар RefSeq-RefSeq). Использовался микрочип Affymetrix GeneChip Human Genome U133 Plus 2.0 Array. Данные по экспрессии генов человека в тканях мозга были взяты из GEO (серия GDS1962), для больных глиомами, в том числе - астроцитомы (6 пациентов - градация опухоли (tumor grade) II, 19 пациентов - градация III), глиобластомы (75 пациентов, градация IV), олигодендроглиомы (37, градация II и 11, градация III).

От общего числа ранговых коэффициентов корреляции Кендалла было рассчитано число статистически значимых для каждого размера выборки. Уровень значимости в зависимости от объема выборки был рассчитан с помощью собственной компьютерной программы и откалиброван с помощью статистических таблиц.

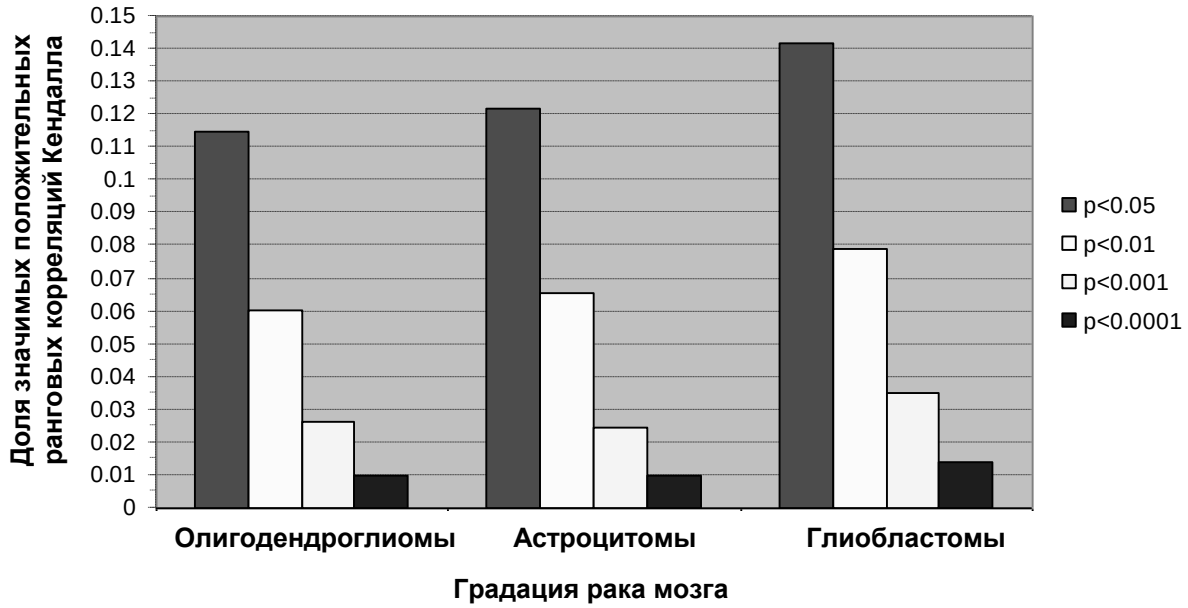


Рис. 2.28. Число значимых позитивных коэффициентов корреляции экспрессии для цис-антисенс транскриптов в геноме человека на выборках глиом мозга (GEO GDS1962).

Доля значимых позитивных коэффициентов ранговой корреляции Кендалла для пар генов, ожидаемое по случайным причинам, не должно превышать уровня статистической значимости, для которого оценивается это число, если пары выбраны случайно. Из рисунка видно, что имеет место положительная корреляция - число, и соответственно доля таких коэффициентов в общем числе коэффициентов, в разы превышает ожидаемое значение. Кроме того, наблюдается увеличение числа таких значимых коэффициентов среди пар при переходе от выборок рака с меньшей градацией агрессивности, к большей (глиобластомы). Таким образом, показан эффект положительно коррелированной экспрессии для цис-антисенс транскриптов в геноме человека на выборках глиом мозга. Феномен корреляции экспрессии транскриптов, находящихся в противоположной ориентации может быть связан с общем увеличением уровня экспрессии генов при раке, когда нарушаются молекулярные механизмы регуляции экспрессии, и она идет постоянно.

В целом, данный раздел представляет компьютерный анализ экспрессии генов с помощью собственных компьютерных программ для микрочипов Affymetrix U133, включающий сложные структуры транскриптов, что является необходимой технической основой дальнейших исследований [48].

2.8. Средства компьютерной интеграции данных

Созданные в ИЦиГ СО РАН методы такой интеграции и верификации данных, ориентированные на платформу SOLiD позволяют сортировать данные по степени достоверности, отсеивать недостоверные результаты и определять место связывания белкового фактора внутри выявленного локуса ChIP-seq [44].

Основная задача компьютерного анализа геномных данных состоит в их функциональной аннотации, интеграции результатов с молекулярно-биологическими информационными ресурсами. В связи с этим большую актуальность приобретает разработка информационно-компьютерных технологий автоматического анализа и функциональной аннотации геномных последовательностей, включая разработку конвейерного подхода (pipe-line) для первичной обработки, процессинга, картирования на референсный геном последовательностей, полученных в ходе масштабного параллельного секвенирования, а также функциональную аннотацию геномных последовательностей с целью разметки регуляторных районов.

Эти направления исследований и технические средства реализованы в разработанном программном комплексе ICGenomics [44]. Особое внимание было уделено оригинальным методам, не повторяющим стандартные алгоритмы для уже достаточно рутинных задач, таких, как выделение кодирующей последовательности или предсказание сайтов связывания транскрипционных факторов (ССТФ) только по нуклеотидной последовательности (с помощью весовых матриц), стандартные решения для которых хорошо проработаны.

Программный комплекс ICGenomics предназначен для компьютерной поддержки исследований в геномике, молекулярной биологии, биотехнологии и биомедицине. Основное назначение – функциональная аннотация геномных последовательностей, получаемых в результате высокопроизводительного секвенирования. Официальное название программы – экспериментальный образец программного комплекса анализа символьных последовательностей геномики (ЭОПК АСПГ).

Программный комплекс ICGenomics позволяет выполнять следующие логически различные функции:

– процессинг (обработку) протяженных последовательностей нуклеотидов из данных секвенирования, полученных с помощью установок секвенирования нового поколения, в том числе: процессинг данных секвенирования платформ 454 и Illumina, процессинг данных секвенирования платформы SOLiD и обработку полногеномных профилей ChIP-seq, включая выделение пиков и предсказание ССТФ;

– аннотацию геномных нуклеотидных последовательностей, включая: разметку положения нуклеосом; поиск экзонов во вновь секвенированных последовательностях; поиск промоторов генов миРНК в нуклеотидных последовательностях на основе специфичных структурных мотивов;

– предсказание аллергенности и функциональных сайтов в пространственных структурах белков; исследование режимов эволюции белок-кодирующих генов.

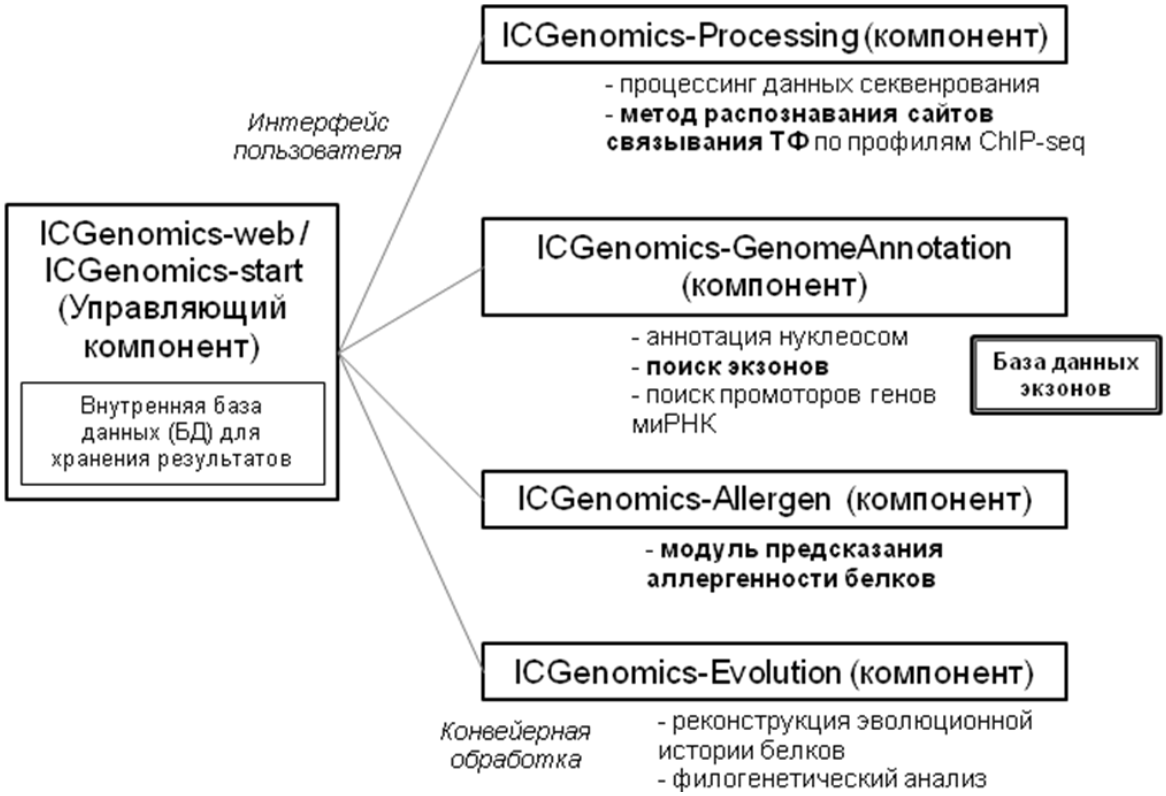


Рис. 2.29. Структура программного комплекса ICGenomics.

Каждая из перечисленных выше функций реализована в соответствующем программном компоненте (рис. 2.29). Программный комплекс состоит из модуля управления (программной компоненты ICGenomics-web и управляющей программы ICGenomics-start) и 4 программных компонент ICGenomics-Processing, ICGenomics-GenomeAnnotation, ICGenomics-Allergen и ICGenomics-Evolution (рис. 2.30).

Входными данными для системы служат файлы нуклеотидных и аминокислотных последовательностей в формате FASTA, а также данные секвенирования в форматах платформ секвенирования Illumina, SOLiD, возможно использование форматов геномных профилей bed (геномных координат), wig (численный профиль). В комплексе используются базы данных SiteEx [72], и PDBSite [512], содержащие скомпилированную ранее информацию об экзонах и пространственных сайтах белков.

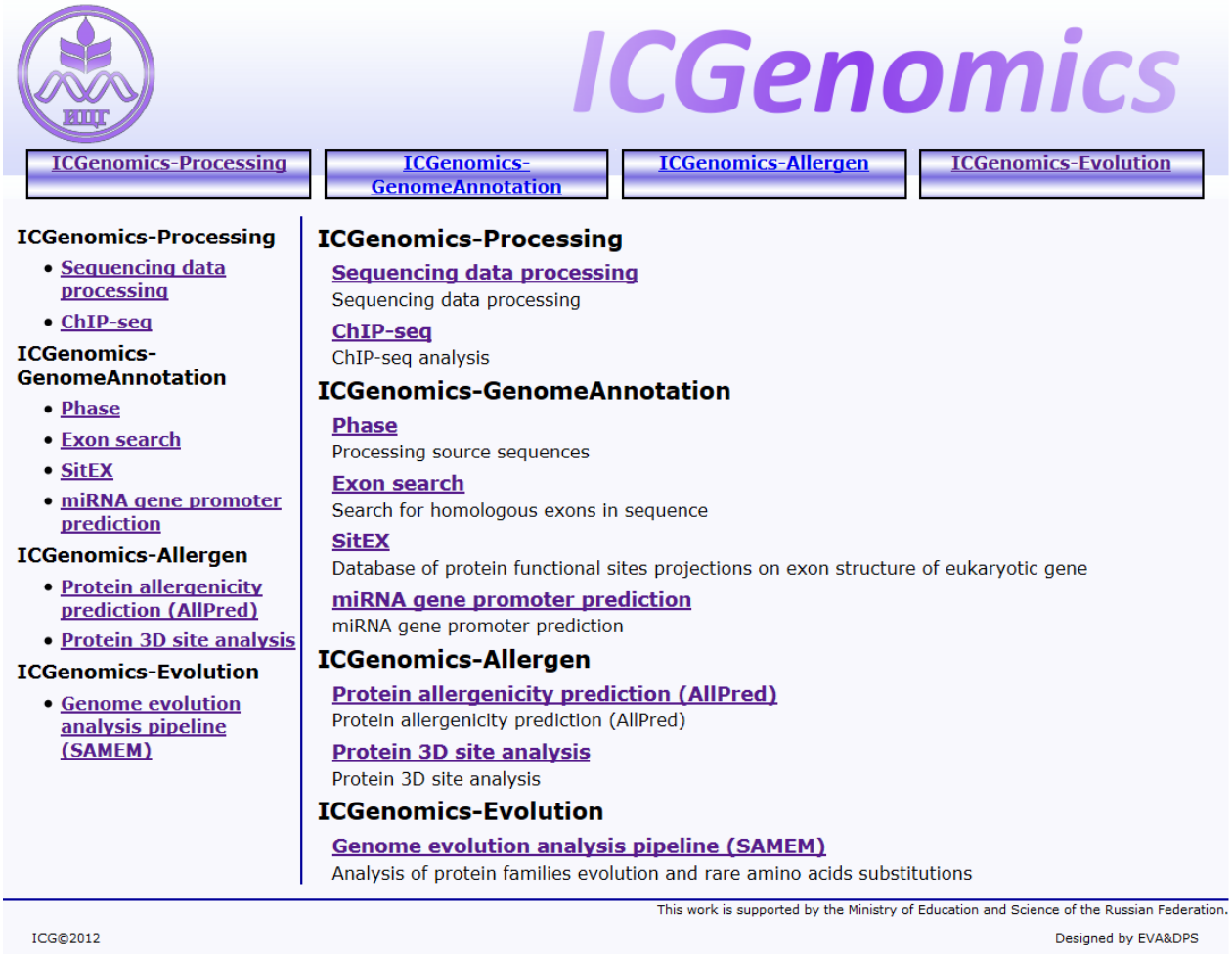


Рис. 2.30. Пример интерфейса управляющего модуля ICGenomics, содержащего функциональные компоненты [44].

Используя тот же формат FASTA, компонент ICGenomics-GenomeAnnotation функциональной аннотации геномных нуклеотидных последовательностей решает смежные задачи функциональной аннотации последовательностей (предсказания позиций нуклеосом, поиска экзонов, поиска промоторов генов миРНК).

Вызов отдельных модулей выполняется из общего интерфейса пошагово. Таким образом, прообразом интеграции средств компьютерной геномики может служить программный комплекс ICGenomics, использующий ряд уникальных модулей. Комплекс применялся к анализу данных ChIP-seq по профилям связывания транскрипционных факторов в геномах мыши и человека [44].

Заключение к Главе 2

Приведенные описывают теоретические компьютерные модели распределения сайтов связывания транскрипционных факторов в геноме по данным ChIP экспериментов. В разделах данной Главы представлены алгоритмы анализа профилей связывания транскрипционных факторов по данным ChIP-seq, выделения пиков, оценки сатурации (полноты) ChIP эксперимента. Показано применение методов для ряда транскрипционных факторов в геноме мыши и в геноме человека. Представлены методы анализа экспрессии генов на микрочипах, оценки качества таких данных [3, 9, 16, 46]. В целом представлены методы полногеномного анализа ChIP-экспериментов, анализа распределения всех сайтов исследуемого транскрипционного фактора в геноме, аннотации их расположения относительно генов.

Материалы настоящей главы обосновывают следующее положение, выносимое на защиту:

Разработанная статистическая модель полногеномного распределения сайтов связывания транскрипционного фактора позволяет оценивать полноту эксперимента по секвенированию и иммунопреципитации хроматина ChIP-seq и рассчитывать статистически значимые оценки нижней и верхней границ общего числа сайтов связывания в геноме для исследуемого фактора.

Представленные результаты позволяют аргументировать следующие выводы:

1) Впервые разработан подход для статистической оценки нижней и верхней границ общего числа сайтов связывания транскрипционных факторов в геноме мыши на основе анализа экспериментальных данных ChIP-seq. Этот подход дает возможность оценки качества экспериментов ChIP-seq для выявления сайтов связывания транскрипционных факторов при заданном объеме секвенирования и размере генома.

2) Разработаны компьютерные методы и программы для анализа данных по полногеномному секвенированию, сопряженному с иммунопреципитацией хроматина, получаемых в экспериментах ChIP-PET и ChIP-seq, и распознавания на этой основе сайтов связывания транскрипционных факторов в геномах человека, мыши, рыбы *Danio rerio*.

Научно-практическая ценность разработанных компьютерных методов, показанных в данной Главе, состоит в возможности поиска регуляторных районов генов по данным секвенирования в масштабе полного генома эукариот. Оригинальные компьютерные программы анализа распределения сайтов связывания

транскрипционных факторов в геноме на основе анализа данных ChIP-seq были представлены в статьях автора [16, 38]. Представлены оригинальные компьютерные программы интеграции данных о сайтах связывания и экспрессии регулируемых ими генов на микрочипах [3, 37, 44]. Разработана база данных качества наборов проб микрочипов платформы Affymetrix U133, и база данных цис-антисенс-транскриптов [38, 48]. Выполнено исследование уровня экспрессии и качества сигнала для наборов проб Affymetrix, таргетные последовательности которых перекрываются в геноме с транспозонами [49]. Создана база данных специализированного микрочипа для генов крысы [53]. Показано объединение программ геномики в программном комплексе ЭО АСПГ, разработанном в ИЦиГ СО РАН [44].

В качестве дискуссии можно привести примеры обобщений разработанных компьютерных подходов анализа данных ChIP-seq на различные геномы эукариот, включая человека, мышь, рыбу *Danio rerio* и дрожжи *S.cerevisiae* [43, 51], подробно описанные в следующих главах настоящей работы.

Глава 3. КАРТЫ САЙТОВ СВЯЗЫВАНИЯ ПО ДАННЫМ ChIP-seq

3.1. Введение. Структура главы

Данная Глава посвящена получению, описанию и анализу полногеномных карт сайтов связывания транскрипционных факторов полученным по экспериментальным данным ChIP-seq в геномах человека и мыши [3, 39, 41, 42, 54]. Решаемые задачи в рамках всей диссертационной работы:

— Компьютерная реконструкция полногеномных карт сайтов связывания транскрипционных факторов c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2 и Smad2 в геноме мыши;

— Реконструкция распределения сайтов связывания транскрипционных факторов MYC, PRDM14, ERα, FOXA1, OCT4, NANOG в геноме человека.

Для экспериментов по секвенированию ДНК, сопряженных с иммунопреципитацией, с помощью методов ChIP-PET и ChIP-seq разработаны компьютерные программы обработки данных, выделения статистически значимых пиков, приведенные в предыдущей Главе. Практическим результатом применения разработанных компьютерных методов определения сайтов связывания транскрипционных факторов в масштабе генома является построение карт связывания с привязкой к хромосомным координатам. С их помощью были проанализированы исходные данные секвенирования и определены сайты связывания белков c-Myc, STAT1, FOXA1, ERα, PRDM14 в геноме человека, транскрипционных факторов и регуляторов Nanog, Oct4, Sox2, Klf4, E2f1, Esrrb, CTCF, n-Myc, c-Myc, Smad1, STAT3, Tcfcf21l, Zfx, Suz12 [3]. Такие карты сайтов связывания используются для дальнейшего полногеномного анализа, определения генов-мишеней, дополнительной экспериментальной проверки (валидации) впервые выявленных регуляторных районов. Карты сайтов связывания важны и для теоретических исследований, кластеризации групп сайтов, реконструкции регуляторных генных сетей.

Первые разделы Главы содержат описание результатов ChIP экспериментов для онкогенов MYC и ERα в геноме человека в культурах клеток рака P493 и MCF-7, соответственно, опубликованных в работах автора [9, 13].

Следующие разделы посвящены массовому ChIP-seq анализу ССТФ в эмбриональных стволовых клетках (ЭСК) мыши [3]. Показано, что число сайтов связывания ТФ в геноме мыши может варьировать от 1 до 40 тысяч, что превосходит

число генов [3]. Выделены кластеры сайтов связывания различных ТФ, построены тепловые карты ко-локализации сайтов в геноме, обсуждена их возможная роль в энхансеоме ЭСК. Для тех же эмбриональных стволовых клеток мыши автором построены карты связывания факторов Eset, Nr5a2 и Smad2. В отдельном разделе представлено исследование сайтов связывания в геноме в зависимости от дозового эффекта (воздействия активатора и ингибитора) на примере Smad2 также в ЭСК мыши.

Выполнено объединение всех сайтов в кластеры в геноме мыши, построена общая карта ко-локализации.

В заключительном разделе приведены результаты анализа ChIP-seq данных для ЭСК человека [42] и показана тот же эффект совместной локализации сайтов факторов плюрипотентности OCT4, SOX2, NANOG, что и для ЭСК мыши.

Исходные данные секвенирования ChIP-seq представлены в GEO NCBI архивами GSE11431 для Nanog, Oct4, STAT3, Smad1, Sox2, Zfx, c-Myc, n-Myc, Klf4, Esrrb, Tcfcp2l1, E2f1, CTCF и регуляторов транскрипции p300 and Suz12, GSE19219 для Tbx3, GSE17439 и GSE17642 для Eset, GSE19019 для Nr5a2, GSE23581 для Smad2 ЭСК мыши, GSE26831 и GSE23893 для ERα человека (культуры клеток) и GSE22767 и GSE22792 для PRDM14 ЭСК человека.

3.2. Распределение сайтов связывания транскрипционного фактора c-Myc, определенное по методу ChIP-PET

Исследование полногеномного распределения сайтов связывания транскрипционных факторов было выполнено для транскрипционного фактора c-Myc в геноме человека на основе технологии ChIP-PET (ChIP - pair-end ditag) (Zeller et al., 2006).

Протоонкоген MYC кодирует транскрипционный фактор c-Myc (здесь и далее также называемый Myc), который регулирует размер клеток, клеточную пролиферацию и апоптоз [68, 323]. Хотя нарушение работы Myc связано с образованием опухолей, неизвестно, какие именно изменения вносит индуцируемое Myc изменение транскриптома в трансформацию клеток. С помощью техники иммунопреципитации хроматина было проведено компьютерное полногеномное исследование прямого связывания фактора Myc с его ДНК-мишенями в модельной культуре В-лимфоидных клеток человека [9].

Митогены индуцируют нормальную экспрессию MYC, когда клетки рекрутируются в клеточный цикл [68], и напротив задержка клеточного цикла (клеточное молчание) и дифференцировка значительно уменьшают экспрессию MYC. В противоположность этому, раковые клетки имеют генетические нарушения регуляции экспрессии гена MYC. Постоянная экспрессия Myc является центральным моментом для их трансформации. Myc – это белок класса «лейциновая застежка» (спираль-поворот-спираль), который димеризуется с белком Max для активации транскрипции [325]. Мотив связывания с ДНК 5’-CACGTG-3’, известен также как E-бокс. Myc также подавляет транскрипцию через взаимодействие с Miz-1 или через другие элементы «корового» промотора [326]; заметим, что механизмы подавления репрессии недостаточно изучены. Транскрипционная активность Myc критична для его способности вызвать опухолевую трансформацию, поскольку аллели с не функциональным транскриптом MYC имеют значительно уменьшенный потенциал трансформации [328]. Большие усилия исследователей были приложены к исследованию роли MYC в образовании опухолей, развитии клеток, направлены на идентификацию генов-мишеней воздействия Myc, и того, как транскрипционные изменения этих мишеней ведут к увеличению размера клеток, прогрессии клеточного цикла, апоптозу, нарушению дифференцировки клеток [330]. Ранее было известно около 1500 генов, показанных как отвечающих на воздействие Myc, и представленных в базе данных его генов-мишеней (www.myccancergene.org) [330]. Высокопроизводительные технологии определения профилей экспрессии в клетке, такие как микрочипы, и серийный анализ экспрессии генов SAGE (serial analysis of gene expression) также применялись для определения сотен генов ответа на Myc. Поскольку большинство исследований экспрессии генов ранее было ограничено возможностями тестирования методом количественной ПЦР (qPCR) определение генов-мишеней этого в масштабе генома было затруднено.

Техника хроматин-иммунопреципитации (ChIP) эффективна для определения прямых генов-мишеней при объединении с методом детекции на микроматрицах (ChIP-chip) или ChIP-qPCR (qPCR продуктов иммунопреципитации) и позволяет определить локусы прямого связывания Myc в геномах [334, 335]. Однако такие методы использования ChIP были сфокусированы только на некоторых специфических характеристиках и участках генома человека. Следовательно, важно было определить прямые гены транскрипционного ответа Myc используя хорошо установленную, интерпретируемую экспериментальную систему и метод, который позволяет

глобальное картирование сайтов связывания Myc. Одним из вариантов решения этой задачи является стратегия полногеномного картирования сайтов связывания транскрипционных факторов ChIP-PET (ChIP в объединении с техникой парных концов PET - Pair-End diTagging) [15] (рисунок 3.1).

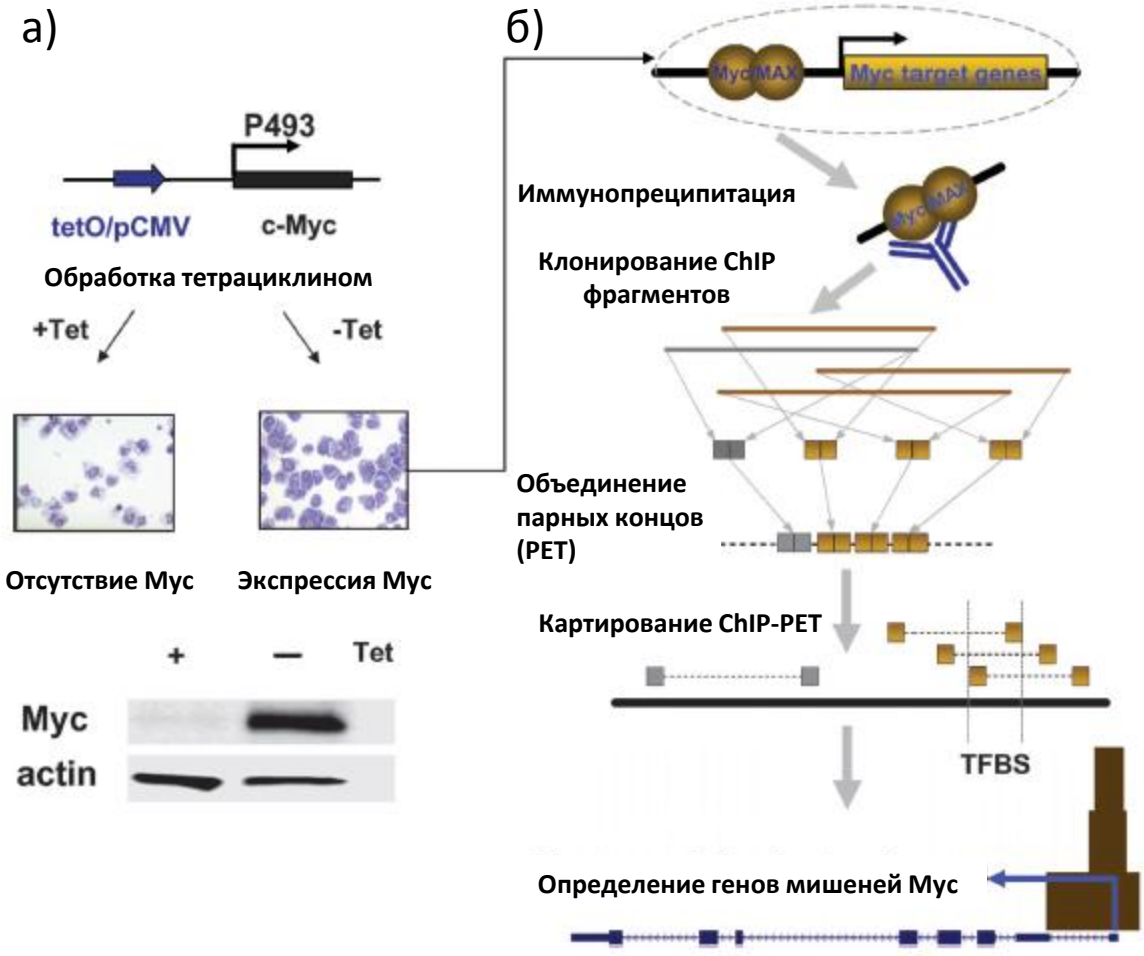


Рис. 3.1. Схема анализа ChIP-PET для определения сайтов связывания Myc в клетках P493 (Zeller et al., 2006).
(а) В-клетки человека, несущие тетрациклин-подавляющую конструкцию c-Myc (линия P493), имеют В-лимфоидный фенотип при культивировании в отсутствии тетрациклина и показывают высокую экспрессию экзогенного MYC при детекции на Вестерн блот.
(б) Схема ChIP эксперимента с использованием поликлонального антитела c-Myc. Парные концы клонированных и выделенных ChIP фрагментов ДНК были объединены (выполнена конкатенация). Парные концы картировались на геном человека для определения связывающих Myc локусов, представленных перекрывающимися кластерами последовательностей (пиками профиля). Показан кластер парных концов размера 3 (PET-3) как пример картирования на первый интрон гена CDK4, известного прямого гена-мишени Myc.

По методу ChIP-PET, выделенные фрагменты иммунопреципитации ДНК сначала клонировались, затем для каждого фрагмента извлекались парные последовательности (тэги) длиной 36 нт с 5’ и 3’ концы. Эти парные тэги (PET) затем

объединялись для эффективного секвенирования и последующего картирования на референсный геном для определения положения выделенных ChIP фрагментов ДНК. Эффективность подхода ChIP-PET была ранее продемонстрирована для картирования сайтов связывания p53 в геноме человека [15] и определения регуляторной транскрипционной сети Oct4 и Nanog в геноме мыши [429]. Для исследования транскрипционной сети ответа на Myc был выполнен эксперимент ChIP-PET на модели В-клеток человека (линия P493). Схема специфичного для Myc анализа ChIP-PET на клеточной линии показана на рисунке 3.1.

Клеточная линия P493 иммортализована посредством введения вируса Эпштейна-Барра и несет тетрациклин-чувствительную конструкцию трансгена MYC. Линия хорошо подходит для глобального картирования сайтов связывания Myc, поскольку имеет почти нормальный кариотип и может формировать у мыши лимфому, подобную лимфому Беркитта [9]. В отсутствии тетрациклина экспоненциально пролиферирующие P493 клетки сверхэкспрессируют Myc и показывают фенотип лимфомы В-клеток.

Связанные с Myc фрагменты ДНК были обогащены с помощью иммунопреципитации хроматина, парные концы длиной 36 нт (две последовательности длиной 18 нуклеотидов с 5’- и с 3’-конца) для каждого ChIP фрагмента извлекались и объединялись в одну последовательность (конкатенировались) для дальнейшего секвенирования и анализа положения в геноме. Последовательности парных концов затем картировались на геном человека для определения границ этих индивидуальных ChIP фрагментов ДНК.

Поскольку сайты связывания Myc обогащены в перемешанном пуле иммунопреципитированной ДНК, множественные уникальные ChIP фрагменты ДНК должны повторять и перекрывать друг друга. Перекрывания положения фрагментов в геномных координатах определяет сайты связывания Myc в геноме. ChIP фрагменты ДНК, которые раздельно расположены в геноме и не перекрываются, скорее всего, неспецифичны. (Рисунок 3.1, панель “б”). В целом в эксперименте на P493 клетках со сверхэкспрессией Myc было сгенерировано около миллиона парных концов (1,143,746 PET фрагментов). Из полученного набора последовательностей PET только 691,966 (61%) однозначно картировались в геноме человека (релиз hg17) и далее были классифицированы как уникальные парные концы, представляющие 273,566 различных обогащенных в ChIP эксперименте фрагментов ДНК. Большинство, 91% из этих последовательностей PET, не перекрывается друг с другом в геноме, и является так

называемыми синглтонами (одиночными фрагментами). Оставшиеся 24,586 фрагмента PET (9%) перекрываются друг с другом, образуя 11,593 кластера парных концов PET, начиная с 2 элементов в кластере до 34. Такие кластеры назовем PET-2 (два перекрывающихся набора PET) и PET-34 (кластер перекрывания 34 PET) (см. Таблицу 3.1).

Таблица 3.1

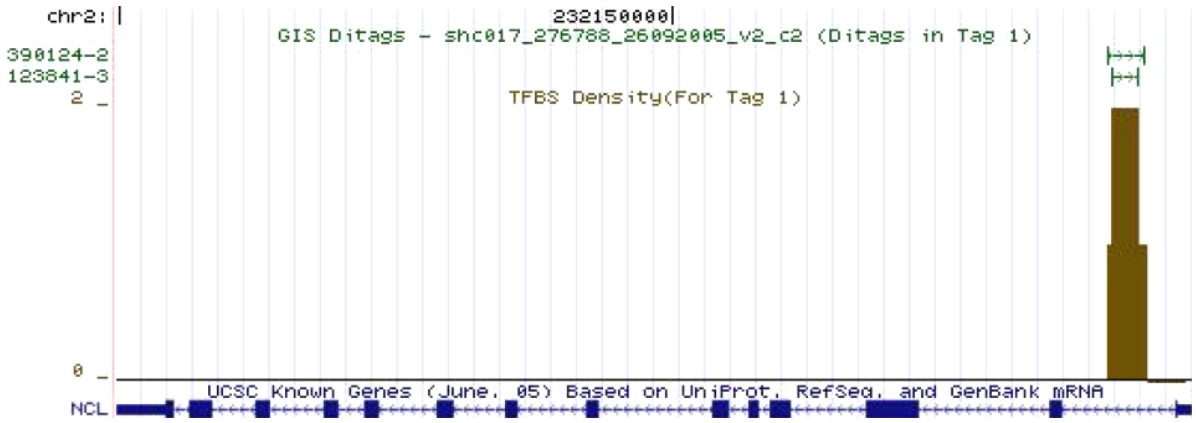
Распределение кластеров ChIP-PET и обогащенность мотивами связывания

ChIP-PET кластеры и мотивы	Компью-терная симуляция	PET синглтоны	PET кластеры			Сайты связыва-ния Myc
			PET-2	PET-3	PET-4+	
Число ChIP-PET фрагментов		248,980	22,500	1,735	351	
Локусы, определенные ChIP-PET	261,948	248,980	11,000	544	49	4,296
Число PET кластеров, которые могут образоваться по случайным причинам		257,251	9,310.9	178.06	2.217	
% шума		100	84.6	32.7	4.5	
Число сайтов с мотивом CACGTG	9,502	10,519	1,329	138	18	1,485
Процент	3.63	4.22	12.08	25.37	36.73	34.60
Число сайтов с мотивом CACATG	57,685	55,342	4,545	263	19	
Процент	22.02	22.23	41.32	48.35	38.78	
Значимость (P-value)		0.00657291	0	0	0.0023266	
Число сайтов с хотя бы одним из мотивов	63,673	62,283	5,194	335	32	2,568
Процент	24.31	25.02	47.22	61.58	65.31	59.80
Значимость (P-value)		0	0	0	~1E-11	

Заметим, что в настоящее время при высокопроизводительном секвенировании максимально могут перекрываться в геноме уже не десятки, а сотни и тысячи прочтений ДНК, образуя пики соответствующей высоты.

Далее для оценки специфичности полученных кластеров PET для определения известного ранее связывания Myc, была проверена локализация PET кластеров относительно генов, известных как прямые гены-мишени Myc *in vivo* [330]. Примечательно, что кластер PET-4 был найден в первом интроне гена NPM1, ранее известного как мишень Myc, кроме того 2 канонических E-бокса были расположены в 86 нт перекрывающемся районе [330]. Более того, 15 других известных гена-мишени

Myc соответствовали кластерам PET-2 и PET-3+, восемь из которых содержали известные сайты связывания Myc. Например, *NPM1*, известный как ген-мишень Myc, имеет кластер PET-2 в первом интроне и два E-бокса внутри 409-нуклеотидного перекрывающегося PET-района; гены *CDK4* (Cyclin-dependent kinase 4), *EIF4E*, *LDHA*, *NME1* (Non-metastatic cells 1, protein NM23A), *PPAT* (Phosphoribosyl pyrophosphate amidotransferase) также имеют кластеры PET2+ [9]. На рисунке 2.3 показано расположение кластера PET-2 связывания Myc в первом интроне гена *NCL*, также известного гена-мишени этого ТФ.



TTTCCCAGCGAGACAACCCCGCCCCTTAAAGAGGCGGCGCGGC**CACGTG**GGGGTACCTGGTAGGCCACCACGTGCCCGAAACAC
AGCGCAAAAGGCACGCGGCAGAGCCACCTTCCGGGGCCCACCACGTGTCGGCTTGCCAGGCCGGGGCGGAGACATCCGAGCGCA
GCCGGGACTCCGACTAGGGCCGATACCGCCATTTCCCGGCCCAAGGCACCCGGTCCCTCTGGAGATTCCAGGACCCACGCGGCG
CGGCC**CACGTG**GCAGGCCGCGCTCCCGGG**CACGT**GCGTCAGGAGTCGAGTCCCAGCTCGTACGCGTCGCAAAAGTTTGGTCTCA
TCTCTCCACCCCGA

Рис. 3.2. Расположение кластера ChIP-PET связывания Myc в интроне гена *NCL* (nucleolin) [9]. Нуклеотидная последовательность района и найденные мотивы связывания CACGTG показаны ниже.

Для оценки уровня ложноположительных предсказаний сайтов в PET кластерах различного размера была выполнена компьютерная симуляция Монте-Карло для расчета расположения возможных кластеров парных концов в геноме человека по случайным причинам. Используя то же число последовательностей той же длины, что и полученные в PET эксперименте, их координаты распределили в геноме (линейной последовательности того же размера, что и референсный геном), используя датчик случайных чисел и собственную компьютерную программу. Рассчитывалось число и распределение по размерам виртуальных кластеров, которые могли бы образоваться при отсутствии специфичности связывания в компьютерном эксперименте.

На основе этой симуляции, вероятность случайного образования 3 и более кластеров фрагментов PET, перекрывающих друг друга в геномных координатах

(кластеры PET-3+), оценена в приблизительно 30%, предполагая, что оставшиеся 70% из 593 кластеров PET-3+ представляют истинное связывание Myc.

Для дальнейшей экспериментальной проверки кластеров PET-3+, ассоциированных со связыванием, было отобрано случайным образом 48 последовательностей кластеров PET-3+ и PET-2, которые тестировались с помощью ChIP-qPCR. Подтверждено, что 100% (29 из 29) кластеров PET-3+ и 47% (9 из 19) кластеров PET-2 действительно были обогащены связыванием Myc (Рисунок 3.3).

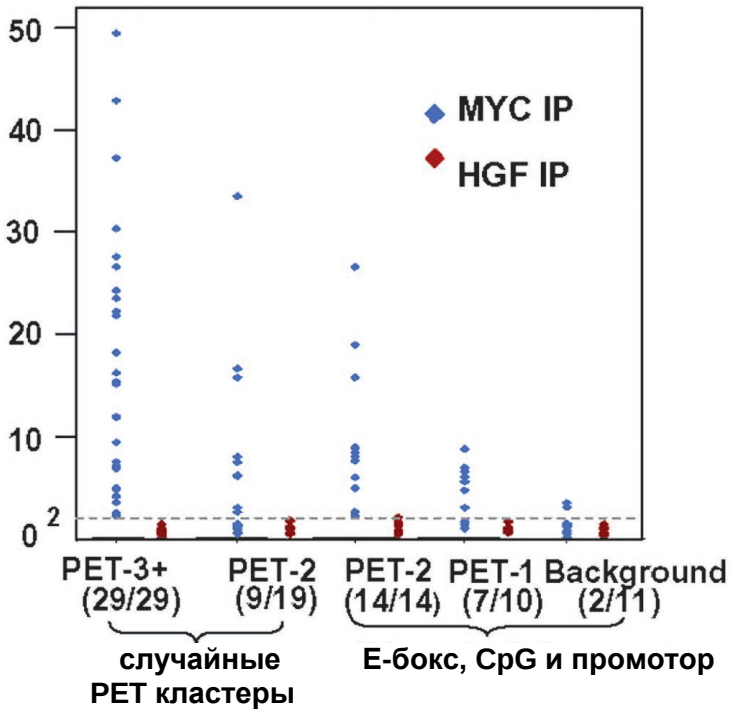


Рис. 3.3. Тестирование специфичности связывания с Myc для PET кластеров разных категорий - выбранных случайно, и кластеров содержащих специфические геномные элементы. ChIP ДНК, связанная с Myc (синий цвет), и контрольная HGF ChIP ДНК (красный цвет) из независимого ChIP эксперимента были проверены на ChIP-qPCR.
Ось Y показывает обогащение связывания (в разы). Первые две полосы слева соответствуют кластерам PET-3+ и PET-2, выбранным случайно. Три полосы справа выбраны из кластеров PET-2, PET-1, и геномных районов, близких к CpG островам и промоторам и содержащих E-бокс. В скобках показано число подтвержденных /тестированных сайтов.

Таким образом, можно заключить, что 593 кластера PET-3+ специфичны. В тоже время связывание Myc подтверждено только для половины из приблизительно 11,000 кластеров PET-2 по результатам статистического анализа и экспериментальной проверки.

Исследование мотива связывания Myc в районах связывания de novo.

Используя 593 экспериментально определенных локуса кластеров PET-3+ связывания Myc, мотив связывания определяли с помощью алгоритма Weeder [230].

Как ожидалось, последовательность CACGTG была наиболее часто встречающимся мотивом, найденным в кластерах PET-3+ (Таблица 2.1). Также определена возможность связывания Myc с неканоническим E-боксом CACATG в этих 593 сайтах высокого качества, подтверждено статистическое обогащение присутствия этого мотива (Таблица 2.1). Более того, в 367 (62%) из 593 участков связывания был найден один или другой из указанных вариантов E-бокса, что соответствует данным работы (12), где была показана высокая аффинность связывания с каноническим и неканоническим E-боксом в клетках, сверхэкспрессирующих Myc. Тем не менее, независимый от E-бокса механизм связывания может быть важным фактором, определяющим сайты связывания Myc, поскольку около 40% исследованных локусов не имеют ни канонического, ни неканонического E-бокса.

Для исследования мотива связывания Myc были использованы только последовательности из надежно определенных участков связывания (кластеров PET-3+, то есть содержащих минимум три перекрывающихся ChIP рида). Известный консенсусный мотив CACGTG (E-бокс) был найден как сверхпредставленный в этих последовательностях. Из найденных 6-меров была получена позиционная весовая матрица, представляющая исследуемый мотив E-бокс (см. рисунок 3.4).

Рис. 3.4. Лого последовательности мотива связывания Myc (E-бокс), полученное по данным ChIP эксперимента с помощью алгоритма Weeder.

Геномные повторы были отфильтрованы («маскированы»), используя аннотацию генома человека в геномном браузере UCSC (версия hg17). Использовалось 593 PET-3+ кластера. Последовательности были обработаны с помощью компьютерной программы поиска мотивов Weeder [230], учитывая следующие параметры: контрольная последовательность для определения частот олигонуклеотидов - геном человека (параметр -HS (Homo sapiens)), поиск в прямой и комплементарной последовательности, допуск множественной встречаемости мотива в одной последовательности, и наиболее полный перебор (тип анализа - “extra”).

Близость сайтов связывания Myc к промоторам и CpG островам. Используя 593 надежных участка связывания Myc, было определено расположение связывания

Мус относительно кодирующих последовательностей генов в геноме человека (рис. 3.5). Большинство участков связывания Мус (63%, 372 из 593) находятся на расстоянии до 10 килобаз вокруг известных районов генов со строгим предпочтением связывания к 5’ проксимальному промоторному району (10 Кб до старта транскрипции и первого интрона).

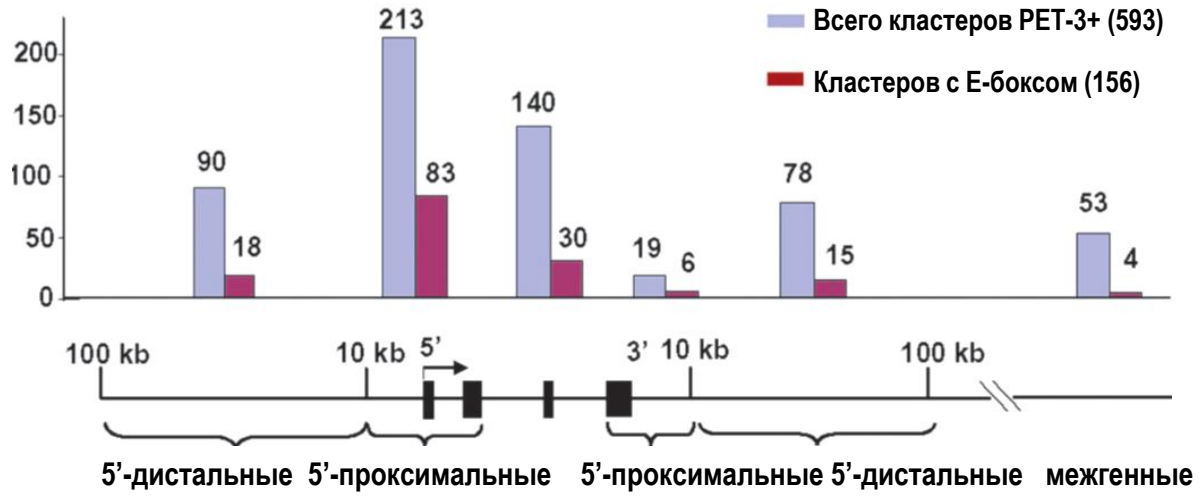


Рис. 3.5. Распределение сайтов связывания Мус относительно структуры гена. Показано расположение 593 надежных участков связывания Мус (синий цвет) и 156 содержащих E-бокс участков (красный цвет) относительно модельной структуры гена - 5’-дистальный, 5’-проксимальный, 3’-проксимальный, 3’-дистальный район и оставшиеся, межгенные участки. Число сайтов связывания показано наверху столбиков схематической гистограммы. Ось Y соответствуют числу сайтов [9].

Примечательно, что Мус связывается в 5’ районе на порядок чаще, чем в 3’ районе (213 против 19 для всех локусов кластеров PET-3+ и 83 против 6 для локусов, содержащих E-бокс). Другая характеристика связывания Мус – это близость к CpG островам, где гипометилированные геномные районы часто ассоциированы с активно транскрибируемыми генами [9]. Проверена ассоциация участков связывания 593 Мус с CpG районами, и найдено, что соответственно 29% и 36% от общего числа кластеров Мус PET-3+ расположены на расстоянии в 1 Кб и 5 Кб от CpG района. Рассматривая подмножество 156 PET-3+ кластеров содержащих E-бокс, мы нашли, что более половины из них (88 участков; 56%) расположены в 5 Кб от CpG.

Эти результаты указывают на предпочтительное связывание Мус с CpG островами в геноме человека. В целом, для более аффинных сайтов связывания (кластеров PET-3+) можно выделить три основные критерия (характеристики) связывания Мус: близость к CpG островам (до 5Кб) и проксимальным промоторным районам генов (в районе до 10 Кб от старта транскрипции и первого интрона),

присутствие E-бокса. Всего 326 из 593 участков связывания Myc определенных кластерами PET-3+ имеют одну из этих трех характеристик. После применения приведенных выше критериев к 11000 потенциальным участкам связывания Myc, предположительно находящимся в кластерах PET-2, было найдено 263 участка, удовлетворяющим всем трем критериям. 1,425 участков соответствовали по меньшей мере двум характеристикам связывания, и 3,703 участка имели по меньшей мере одну из характеристик связывания Myc (1,320 содержали E-бокс, 1,854 были в интервале 5Кб от CpG островов и 2,212 находились в проксимальных промоторных районах).

Чтобы оценить экспериментально истинный уровень связывания Myc к участкам кластеров с различными характеристиками, было выбрано случайным образом 10 участков из кластеров PET-2 с одной из трех, двумя из трех или всеми тремя характеристиками связывания Myc для проверки методом ChIP-qPCR. Как ожидалось, все 14 участков (100%), удовлетворяющие всем трем критериям, показали связывания (см. Рис. 3.3). Процент подтвержденных участков связывания из других шести категорий варьировал от 20% (расположение в 10 Кб от промоторов) до 58% (присутствие E-бокса и близость к CpG островам). Среди всех характеристик участки, содержащие E-бокс, показали наивысший уровень подтверждения (Рис. 3.3). Эти результаты позволяют предположить, что менее 50% из имеющихся 11000 последовательностей кластеров PET-2 являются истинными сайтами связывания Myc.

Далее была предпринята попытка оценить исходя из имеющегося распределения кластеров по размерам, сколько сайтов связывания Myc были пропущены кластерами PET или попали в одиночные участки PET (синглтоны). Предполагаемые участки связывания, которые удовлетворяли всем трем характеристикам связывания Myc, были определены в PET-синглтонах (587 участков) или участках генома, не попавших в кластеры (всего 3,689 районов), и протестированы с помощью ChIP-qPCR. Семь из 10 сайтов PET-1 (70%) и 2 из 11 (18%) участков геномного фона (не занятых фрагментами PET в проведенном эксперименте) подтвердили связывание, но на невысоком уровне сигнала qPCR (Рис. 3.3), предполагая, что много потенциальных сайтов низкой аффинности были не детектированы в эксперименте кластерами PET из-за недостаточно глубокого секвенирования. В целом, для пополнения данных с учетом более слабых взаимодействий Myc–ДНК, 3,703 участка кластеров PET-2, содержащих по меньшей мере одну из характеристик, связывания были скомбинированы с 593 сайтами полученными из кластеров PET-3+, определяя таким образом общее число 4,296 сайтов связывания Myc в геноме, найденных в проведенном ChIP-PET

эксперименте. Поскольку эти 4,296 сайтов потенциально имеют 50% ложноположительных сайтов, далее они рассматривались с учетом дополнительной характеристики - изменения экспрессии близлежащих генов.

Гены-мишени воздействия транскрипционного фактора могут быть определены по расположению сайтов связывания относительно границ гена в геноме. Задав расстояние близости к гену в 10Кб, и сравнивание расположение сайтов связывания Myc с полным списком всех аннотированных генов в геноме человека (аннотация UCSC Known genes) получаем набор потенциальных генов-мишеней (генов - кандидатов прямого воздействия). Описанные выше 4,296 сайтов связывания Myc дают по такой процедуре список в 2,980 генов. Если увеличить расстояние до 100Кб, получим уже 3,465 генов, ассоциированных с сайтами связывания Myc. По использовавшейся аннотации «UCSC Known genes», геном человека содержит 25 тысяч генов, таким образом, доля генов прямого воздействия Myc в B-клетках составляют 12-14%.

Чтобы проверить, какие из этих генов отвечают на активацию Myc, использовались данные измерения экспрессии генов на микрочипах платформы Affymetrix U133 в той же клеточной линии (P493) с обработкой и без обработки тетрациклином. Был получен набор дифференциально экспрессирующихся генов по данным микрочипов, который затем сравнивался с наборами генов-кандидатов прямого воздействия Myc по расположению сайтов связывания. Из 3,465 предполагаемых генов прямого воздействия Myc по расположению сайтов, 668 генов дифференциально экспрессировались на микрочипах при использовании порогового значения уровня значимости микрочипов q<.05% [153]. Из этих 668 генов, отвечающих на воздействие Myc, 406 генов значимо повышали экспрессию и 262 понижали экспрессию.

Функциональная классификация этих генов-мишеней Myc была выполнена с помощью категоризации генных онтологий (ГО) на основе базы данных PANTHER (http://panther.appliedbiosystems.com). Было выделено 211 различных функциональных категорий. Для этих 668 генов многие категории онтологий из таких групп как метаболизм (metabolism), контроль клеточного цикла (cell cycle control), регуляция транскрипции (transcription regulation), каскад внутриклеточных сигналов (intracellular signal cascade), и биосинтез (biosynthesis) статистически обогащены (сверхпредставлены) на уровне (P < 0.05) с учетом статистической поправки на множественность гипотез. Такой набор категорий генных онтологий соответствует представлению о том, что Myc воздействует на общие регуляторные генные сети со

специфическим влиянием на метаболизм, увеличение размера клетки и клеточную пролиферацию [9].

Мус воздействует на гены, регулирующие транскрипцию, образуя регуляторные контуры. Функциональная классификация генных онтологий этих 668 прямых мишеней Мус дает группу генов метаболизма нуклеиновых кислот (140 из 668; 21%) как наибольший статистически значимый класс, из них 49 генов кодируют транскрипционные регуляторы. Среди генов транскрипционных регуляторов, напрямую активируемых Мус (повышающих экспрессию), MAX, MXI1, MXD3, и MNT, которые вовлечены в сеть белковых взаимодействий Мус-Max-Mad. Это наблюдение предполагает дополнительный уровень регуляции в этой белковой сети, в которой члены семейств белков Мус и Mad формируют гетеродимеры с Max. Были найдены и другие факторы, связанные с контролем роста клетки и регуляцией клеточного цикла, такие как NFκB, STAT3, ERa, JUN, ELK-4, CEBP и ETS1.

Прямые гены-мишени репрессируемые Мус. Несмотря на то, что по данным микрочипового анализа многие гены были выделены как репрессируемые Мус (значимо понижающие экспрессию), только несколько генов-мишеней, таких как CDKN2B и CDKN1A, были ранее описаны как прямые мишени репрессии Мус [335]. Полногеномное картирование определяет группу 262 генов, репрессируемых Мус, которые одновременно связаны Мус (на заданном расстоянии). Из функциональной классификации генных онтологий в этих прямых генах-мишенях репрессии Мус статистически обогащены каскад внутриклеточных сигналов (intracellular signaling cascade), передача сигнала (signal transduction) и метаболический путь созревания В-клеток. Для групп сайтов связывания Мус, соотнесенных с активаций и репрессией по дифференциальной экспрессии их генов-мишеней, был выполнен анализ нуклеотидных последовательностей и поиск нуклеотидных мотивов. Использование весовых матриц TRANSFAC показало, что два типа сайтов связывания транскрипционных факторов EBF (early B cell factor - фактор ранних В-клеток) (P<4.48E-19) и ZIN3 (P<5.52E-15), статистически обогащены в наборе генов, репрессируемых Мус, по сравнению с генами, индуцируемыми Мус. Транскрипционный фактор EBF необходим для образования и спецификации В-клеток.

Кофакторы, действующие совместно с Мус в цис-регуляторных модулях. Известно, что кроме облигатного партнера связывания Max, фактор Мус взаимодействует с другими транскрипционными комплексами, и активация генов модулируется через эти взаимодействия воздействием Мус [327]. Например, Мус

формирует комплекс с Miz-1 для подавления экспрессии генов [68]. В отдельных случаях Myc может взаимодействовать с AP-2, C-EBP, HIF-1, Sp1 или Sp3. Тем не менее, не было точно определено какие другие транскрипционные факторы ко-регулируют их гены мишени совместно с Myc. Вопрос анализа цис-регуляторных модулей, отвечающих на Myc совместно с другими транскрипционными факторами, был изучен на последовательностях сайтов связывания из наиболее надежных 593 сайтов Myc, заданных кластерами PET-3+, с помощью весовых матриц базы данных TRANSFAC. Использовались оптимизированные процентные весовые матрицы для более чем тысячи сайтов связывания (1,051) транскрипционных факторов человека из TRANSFAC (версия 9.1) [129].

Мотивы связывания 20 различных ТФ были значимо обогащены ($P<10^{-20}$) с как минимум троекратным превышением ожидаемой частоты встречаемости (от 3- до 24 раз) по сравнению с контрольными геномными последовательностями. Среди этих мотивов транскрипционных факторов, мотив связывания Myc:Max был перепредставлен более чем в 10 раз ($P<10^{-180}$). Были найдены и другие мотивы известных партнеров Myc, такие как AP2 и Sp1. Анализ функциональных категорий генных онтологий генов, ассоциированных с этими консенсусными сайтами связывания транскрипционных факторов, позволяет предположить ассоциации специфических генных функций со связыванием Myc и Sp1, AP-2, или MAZ. Мотив E2F1 специфически обогащен (в 16 раз) в ChIP кластерах связывания Myc, и в 37 раз обогащен в подмножестве кластеров, содержащих мотив E-бокс. При пересечении набора генов с данными экспрессии на микрочипах, 67 из 171 идентифицированного локуса были ассоциированы с генами, чья экспрессия была модулирована в клетках P493 (52 гена имели повышенную экспрессию и 15 - пониженную). Среди них гены CDC6, DHODH, MCM3, и MCM4 были подтверждены как связанные и индуцируемые обоими транскрипционными факторами, E2F1 и Myc. Подобно Myc, E2F1 также контролирует прогрессию клеточного цикла и репликацию ДНК. Таким образом, де-регулирование Myc потенциально может привести к неконтролируемой прогрессии клеточного цикла через функциональную связь с E2F1 [9].

Функции сайтов связывания Myc и компьютерные оценки распределения числа сайтов Myc в геноме. Для оценки максимального числа уникальных ChIP фрагментов ДНК, полученных в исходной клонированной ChIP библиотеки, был выполнен анализ полноты эксперимента (сатурации). Из общего числа 691,966 PET фрагментов, картированных на релиз генома человека hg17, было определено 273,566

различных PET элемента (парных последовательностей). Распределение этих 273,566 последовательностей, сгруппированных в кластеры, можно описать экспоненциальной функцией распределения, где 38% (105,520) из 273,566 последовательностей PET представлены одиночными копиями (синглтоны), и только малая доля - 0.93% (2,568) всего набора картированных последовательностей представлены 10-24 копиями. Среднее значение - 2.55 PET последовательности на уникальный участок генома (пик профиля). Можно предположить, что с увеличением числа PET последовательностей в эксперименте будет увеличиваться уровень покрытия сайтов кластерами фрагментов ДНК и качество определения сайтов в геноме, что и было подтверждено впоследствии многочисленными экспериментами ChIP-seq. Общее число найденных участков в PET кластерах должно «насыщаться», образуя плато (стабильное состояние) в распределение числа фрагментов при таком экспериментальном пополнении, как описано в предыдущей Главе. Действительно, число синглтонов уменьшалось, а число PET кластеров большего размера - увеличивалось [9].

Было выполнено сравнение наблюдаемого распределения размеров PET кластеров с компьютерной моделью распределения таких кластеров. Использование фрагментов PET той же длины, что и в эксперименте – при этом положение кластера на хромосоме генерировалось случайно, а длина выбиралась случайно из эмпирического распределения длин. Рисунок 3.6 представляет сравнение числа ChIP-PET кластеров для MYC в геноме в зависимости от числа перекрывающихся фрагментов в кластере, для 273 тысяч пар PET.

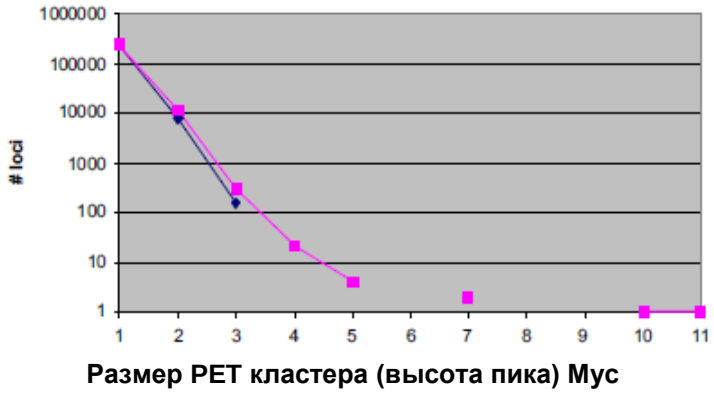


Рис. 3.6. Распределение числа ChIP-PET кластеров для MYC в геноме в зависимости от числа перекрывающихся фрагментов в кластере. Представлено наблюдаемое распределение (верхняя линия) и смоделированное распределение (нижняя линия) случайно расположенных фрагментов [9].

Для оценки чувствительности метода ChIP-PET и числа прочтений последовательности, необходимых для детектирования всех потенциальных сайтов

связывания Myc в геноме P493 использовался анализ подгонки кривых распределения числа кластеров PET прочтений ДНК в эксперименте. Показано, что только около 6% всех прочтений ДНК в ChIP библиотеке были получены специфической иммунопреципитацией, остальные представляли собой шумовой сигнал эксперимента (Zeller et al., 2006). Была выполнена аппроксимация распределения PET кластеров последовательностей, полученных в эксперименте, специфически иммунопреципитированных с белком, с помощью функции Парето [16], используя данные ChIP-qPCR (свыше 200 испытаний). Распределение аффинности связанных с Myc фрагментов ДНК сильно скошено с большинством сайтов, имеющих относительно низкую аффинность, только несколькими сайтами, имеющими высокую аффинность. Применяя такое распределение аффинности к данным ChIP-PET эксперимента - распределению кластеров фрагментов по силе связывания - общая оценка числа сайтов в геноме P493 достигает 20000 сайтов. Отметим, что с увеличением глубины секвенирования в рамках предложенной модели будет увеличиваться доля кластеров больших размеров (3 фрагмента и выше, 4 фрагмента и выше, и т.д.) в сайтах связывания в геноме.

Заключение к разделу

Связывание Myc потенциально занимает более 4000 локусов в геноме, большинство из которых проксимальные промоторные районы, часто ассоциированные с CpG островами. С использованием профилей экспрессии генов на микрочипах вместе с ChIP-PET были определены 668 непосредственных генов-мишеней, регулируемых Myc, включая 48 транскрипционных факторов. Таким образом, показано, что Myc является транскрипционным узлом в контроле роста и пролиферации клеток. Это первое экспериментально полученное полногеномное распределение сайтов связывания Myc дает представление о транскрипционных контурах и цис-регуляторных модулях, вовлекающих Myc, задает стандарт анализа геномных механизмов образования опухолей, вызываемых Myc [9].

3.3. Исследование распределения сайтов связывания ТФ рецептора эстрогенов ERα с помощью ChIP-seq

Проведено компьютерное исследование сайтов связывания ТФ ERα (эстроген-рецептор) в геноме человека по данным ChIP-seq в культурах раковых клеток MCF-7 и

T47D [13]. Из огромного набора потенциальных сайтов связывания для ТФ, определенном только по последовательности букв 13 нуклеотидов матрицей связывания, лишь малая часть связывается с ТФ in vivo (1-2%), как было показано с помощью компьютерных оценок автора в работе [13]. Для рецептора эстрогенов – число сайтов в геноме человека – около 1 миллиона по консенсусу (с несовпадениями), ~32 тыс. сайтов по позиционной весовой матрице, и число пиков, экспериментально определенных сайтов с помощью ChIP-seq составляет около 17 тысяч.

Интерес представляет определения «активных», или реально связанных сайтов в геноме на основе контекстных характеристик, других экспериментальных данных или особенностей генома. С помощью анализа данных ChIP-seq был определен нуклеотидный мотив и уточнена весовая матрица связывания (лого) (Рис. 3.7).

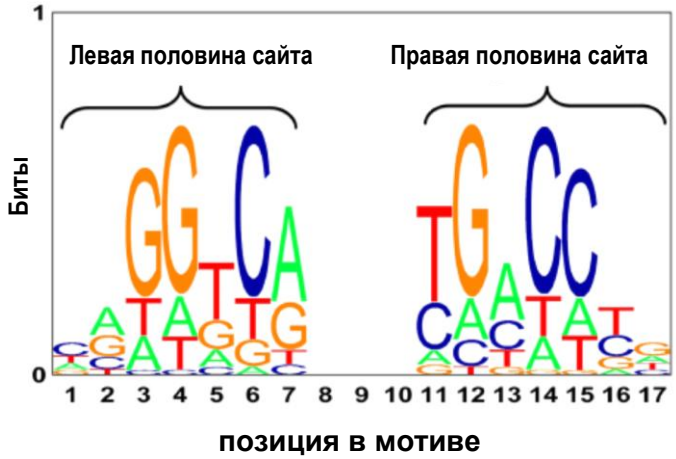


Рис. 3.7. Определение палиндромного мотива связывания de novo для ССТФ ER в геноме человека [13].

С помощью разработанных автором методов были определены гены-мишени и оценено изменение уровня экспрессии этих генов, измеренное на микрочипах платформы Affymetrix в культуре клеток MCF-7. После обработки культуры клеток эстрадиолом (E2) экспрессия гена увеличивается в разы (по сравнению с нейтральным состоянием, когда культура клеток в растворе, в этаноле).

Показана корреляция между силой связывания измеренной для 81 сайта ER и высотой пика геномного профиля ChIP-seq в логарифмической шкале (рис. 3.8).

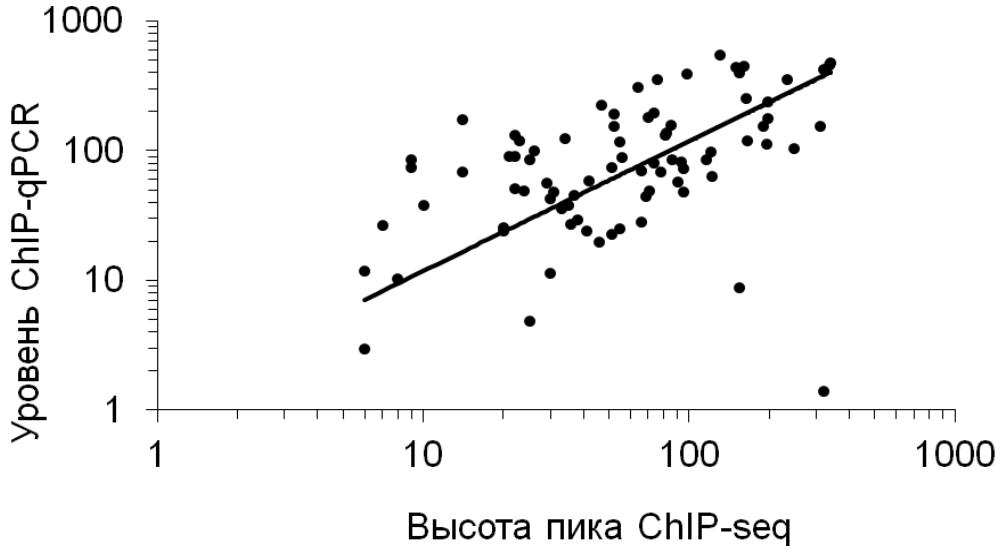


Рис. 3.8. Экспериментальная проверка 81 сайта связывания ER с помощью ChIP-qPCR. Выборка из сайтов связывания ER определенных ChIP-seq [13].

Линейный коэффициент корреляции (CC) составил 0.56 (P=5.0E-8), ранговый коэффициент корреляции Кендалла 0.375 (P= 7.17E-07). Высокая корреляция еще раз подтверждает адекватность модели определения связывания транскрипционного фактора с помощью эксперимента ChIP-seq.

Для детального исследования и определения параметров, определяющих связывание транскрипционного фактора с последовательностями ДНК с подходящим мотивом, был использован рецептор эстрогенов альфа ERα (здесь и далее может упоминаться как ER) в качестве модельной системы.

За основу анализа связывания ER был взят набор сайтов, определенных в ChIP-seq эксперименте в клеточной линии MCF-7 после обработки эстрадиолом (E2). Используя описанный ранее алгоритм (Глава 2), было определено около 17 тысяч сайтов. Поскольку известно, что линия MCF-7 содержит хромосомные перестройки и протяженные амплифицированные участки хромосом (существующие в двух и более копиях), для статистического анализа было решено не использовать сайты связывания, лежащие в амплифицированных участках генома MCF-7 [513]. Таблица таких участков генома MCF-7, содержащих значительные отклонения в копийности, приведена в Приложении. Оставшееся число сайтов связывания ER в клетках MCF-7 оставило 16043.

Рассмотрим воспроизводимось (репродуцируемость) экспериментов, использующих иммунопреципитацию хроматина, сравнивая число сайтов в геноме человека, найденных в каждом опубликованном эксперименте.

Использовались данные [13], 1,226 сайтов, найденных с помощью ChIP-PET [346], 8,225 сайтов связывания из ChIP-on-chip эксперимента [339] (более полная версия по сравнению с ChIP-chip данными о 3,665 сайтах; Carroll et al, 2006) и 10,191 сайтов, идентифицированных с помощью ChIP-seq [347]. Перекрывание с набором сайтов [13] составляет 69%, 74% и 62%, соответственно.

Перекрывание наборов сайтов рассчитывалось по перекрыванию геномных координат сайтов, допускалась разница не более 200 нт между сайтами из сравниваемых наборов (200 нт - это точность ChIP-seq эксперимента, связанная со средней длиной фрагмента секвенирования).

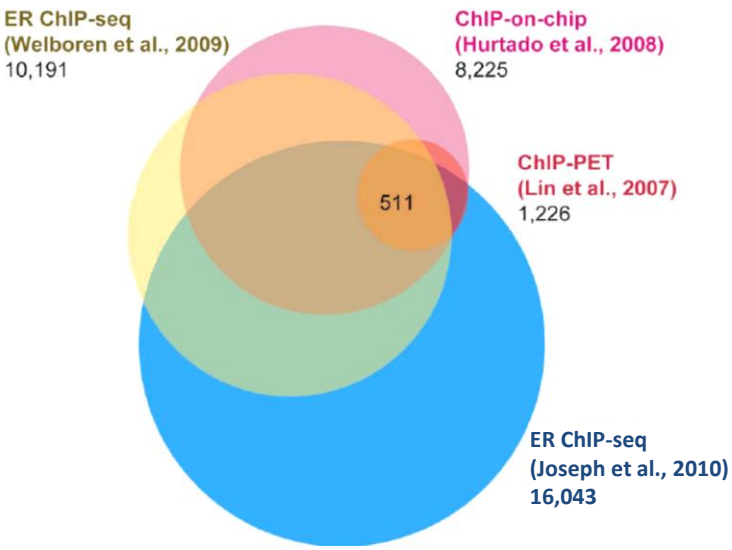


Рис. 3.9. Диаграмма Венна воспроизводимости определения сайтов связывания ER используя ChIP технологии в геноме человека: перекрывание наборов сайтов, представленных в опубликованных ранее статьях в том же типе клеток MCF-7. Перекрывание составляет 69%, 74% и 62%, соответственно.

Точные числа совпадения показаны в таблице. Матрица перекрывания между наборами сайтов не симметрична, поскольку один сайт, в рассматриваемом эксперименте может содержать два расположенных рядом сайта из другого набора данных. Это связано как с технологическими особенностями (кластеры прочтений ChIP-PET шире, чем пики профиля ChIP-seq), так и с молекулярно-биологическими

характеристиками (два расположенных рядом сайта, например в промоторе, неразличимы и дают один сигнал в другом эксперименте).

Таблица 3.2

Пересечение числа сайтов ER, найденных в эксперименте ChIP-seq [13] с опубликованными ChIP-chip, ChIP-PET и ChIP-seq

	ER-ChIP-seq (Joseph *et al*., 2010)	ChIP-PET (Lin *et al*, 2007)	ChIP-chip (Carroll *et al*, 2006)	ER-ChIP-seq (Welboren *et al*, 2009)
Общее число сайтов	16,043	1,226	3,665	10,191
ER-ChIP-seq (Joseph *et al*., 2010)	-	841	3,152	5,823
ChIP-PET (Lin *et al*, 2007)	838	-	610	814
ChIP-chip (Carroll *et al*, 2006)	2,977	608	-	2,041
ER-ChIP-seq (Welboren *et al*, 2009)	6,291	955	2,590	-

Рисунок 3.10 показывает распределение силы связывания сайтов связывания ER в геноме человека, измеренной с помощью qPCR в клетках линий MCF-7 (левая панель) и T47D (правая панель).

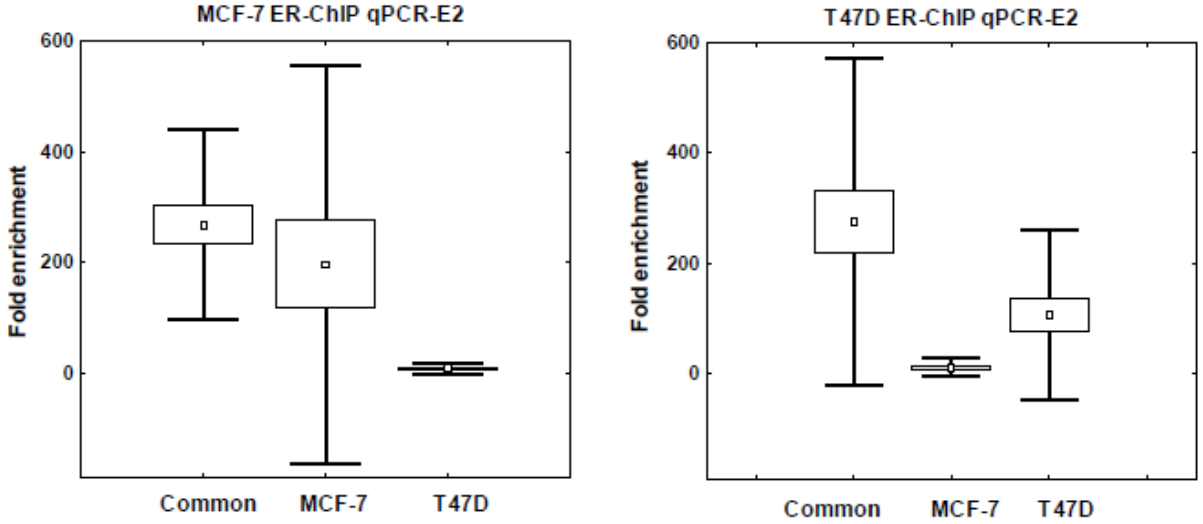


Рис. 3.10. Распределение аффинности сайтов связывания ER в геноме человека, измеренной с помощью qPCR в клетках линий MCF-7 (левая панель) и T47D (правая панель).

Сайты, специфичные для линии MCF-7 (не имеющие выраженного пика ChIP-seq в T47D), действительно имеют более сильное связывания в qPCR эксперименте в клетках, и наоборот, сайты специфичные для линии T47D по ChIP-seq эксперименту имеют более сильное связывание по qPCR в клетках T47D. Отметим, что сайты, общие

для двух клеточных линий, показывают наиболее высокий уровень связывания по qPCR.

Была рассчитана корреляция между ChIP-seq связыванием сайтов ER до и после обработки эстрадиолом (E2) в клетках линии MCF-7. Коэффициент линейной корреляции r=0.662 (уровень значимости P < 2.2e-16), коэффициент ранговой корреляции Кендалла τ=0.328 (P< 2.2e-16). В целом наблюдается два эффекта: корреляция связывания ER в двух типах клеток и появление новых сайтов в клетках линии MCF-7, имевших незначимый (практически нулевой) сигнал в клетках T47D.

Позиции сайтов связывания ER отобранных в ChIP-seq эксперименте после активации эстрадиолом были проверены в независимом эксперименте ChIP-seq на связывание в клетках без обработки (растворитель DMSO). Сила связывания представлена высотой пика в двух независимых экспериментах ChIP-seq. Высота пика профиля ChIP-seq в среднем много выше для сайтов связывания после обработки клеток эстрадиолом (E2). Коэффициент линейной корреляции r=0.662 (уровень значимости P < 2.2e-16), коэффициент ранговой корреляции Кендалла τ=0.328 (P< 2.2e-16).

Поскольку ТФ FOXA1 был предложен в качестве так называемого «первооткрывающего» фактора (pioneering factor), который потенциально может направлять связывание ERα, [319, 320] было проведено следующее исследование. Было рассчитано, насколько присутствие связывания FOXA1, измеренное с помощью технологии ChIP-seq в той же культуре клеток, в отсутствии обработки эстрадиолом может быть предсказывающим маркером связывания ERα уже после обработки клеток этим лигандом. Известно также, что мотив фактора AP1 является ко-мотивом в сайтах связывания ER. Поэтому два компонента комплекса AP1, а именно FOS и JUN были также включены в ChIP-seq анализ. Детали примененных антител для ChIP-seq экспериментов, и параметры библиотек секвенирования даны в таблице 3.3 и в работе автора (Joseph et al., 2010).

Если ранжировать все выделенные сайты по высоте пика и разделить этот список на 4 группы (квартили, 25% от общего числа генов в списке, что составляет около 4 тысяч сайтов), можно качественно исследовать поведение профилей ассоциации с другими маркерами в геноме. Такое разделение позволяет грубо оценить вклад сайтов с различной силой связывания в общую картину среднего распределения профилей маркеров хроматина, фазированных относительно центральной позиции сайта.

Таблица 4.4 показывает разделение исследуемого набора сайтов связывания ER на квартили, процент содержания мотива связывания ER, определенного по весовой матрице и число сайтов ER, содержащих сайты связывания транскрипционного фактора FoxA1.

Таблица 3.3

Процент и число сайтов содержащих мотив FOXA1 по квартилям сайтов ER

Квартиль набора сайтов связывания ER	Число сайтов связывания ER	Процент содержания мотива связывания ER (%)	Число сайтов связывания ER содержащих сайты связывания FOXA1
Q1	4360	82.1	3581
Q2	4360	43.1	1877
Q3	4360	24.1	1052
Q4	4362	15.3	668

Из таблицы видно, что наибольший процент содержания мотив связывания - в первой квартили. Наибольшее число сайтов связывания FOXA1 также присутствует в первой квартили набора сайтов ER, что свидетельствует о совместной работе этих факторов.

Следующая таблица 3.4 детализирует представленность мотива связывания ER по совпадению с консенсусной последовательностью GGTCAnnnTGACC; одно несовпадение («mismatch») обозначено mm1, два несовпадения - mm2, полное совпадение (нет несовпадений) - mm0.

Таблица 3.4

Присутствие консенсуса связывания и сайтов FOXA1 в сайтах связывания ER по квартилям

Сайтов ER с FoxA1 по квартилям	Полное совпадение mm0	Одно не-совпадение mm1	Два не-совпадения mm2	Нет мотива связывания	Всего сайтов
Q1 и FoxA1	143	674	1387	1377	3581
Q2 и FoxA1	42	195	635	1005	1877
Q3 и FoxA1	5	92	294	661	1052
Q4 и FoxA1	1	26	171	470	668
Все сайты ER					
Q1	192	803	1707	1658	4360
Q2	134	637	1578	2011	4360
Q3	102	591	1482	2185	4360
Q4	84	475	1460	2343	4362

Из таблицы видно, что первая квартиль ранжированного по силе связывания набора сайтов связывания ER содержит большее число совершенных совпадений с консенсусом связывания и большее число сайтов FOXA1. Наблюдается градиент от первой квартили к четвертой по выраженности мотива связывания. В то же время присутствие сайтов FoxA1 частично компенсирует отсутствие четкого мотива связывания ER. По-видимому, этот кофактор облегчает открытие хроматина и связывания белка ER со своими сайтами.

Рассмотрим вопрос о воспроизводимости ChIP-seq эксперимента по определению сайтов связывания ER в целом и по квартилям.

Следующая таблица 3.5 представляет результаты пересечения ChIP-seq (Welboren *et al*, 2009), ChIP-chip (Carroll *et al*, 2006) и ChIP-PET (Lin *et al*, 2007) экспериментов по квартилям.

Из таблицы видно, что наибольшую воспроизводимость по пересечению с опубликованными ранее наборами данных имеет первая квартиль сайтов. Наблюдается градиент по воспроизводимости - самые «сильные» сайты связывания ER лучше воспроизводятся в других экспериментах, также основанных на иммунопреципитации хроматина.

Таблица 3.5

Воспроизводимость ChIP эксперимента по определению связывания ER в клеточной линии MCF-7

Квартили сайтов связывания ER	ChIP-PET (Lin *et al*, 2007)	ChIP-chip (Carroll *et al*, 2006)	ChIP-seq (Welboren *et al*, 2009)	Пересечение по всем наборам
Q1	626	1,945	2,749	407
Q2	129	727	1,497	32
Q3	56	319	969	6
Q4	30	161	608	1
Всего пересечений	841	3,152	5,823	446
Всего сайтов	1,226	3,665	10,191	-

Был выполнен более детальный анализ мотива связывания ER, используя расчет для каждой половины димера GGTCAnnnTGACC. Поскольку матрица связывания симметрична, а мотив связывания в целом представляет собой палиндром, можно определить только «полусайт» - половину сайта связывания, а затем уже искать вторую часть, если она вырождена. Была построена оптимизированная энергетическая матрица связывания используя алгоритм TherMoS (Thermodynamic Modeling of chip-Seq). Кратко, алгоритм состоит в итеративном уточнении матрицы связывания в нуклеотидных последовательностях заданных пиков ChIP-seq, используя в качестве меры оптимизации высоту пика и вероятность наблюдения прочтений ДНК в окрестности пика. Детальное описание алгоритма и применение на более широком наборе ССТФ описано в работе (Sun et al., 2013). При обучении алгоритма на ChIP-seq профиле ER на участках размером 1Кб, центрированных на 16,043 пиках связывания (+/-500 нт), алгоритм определил палиндромный мотив, показанный на рисунке 3.7.

При генерации рисунка, позиционно-специфичная матрица энергии связывания - PSEM (Position Specific Energy Matrix) была конвертирована в традиционную позиционно-специфичную частотную матрицу (весовую матрицу), используя экспоненциальную трансформацию [220]. Мотив, показанный на рисунке, построен как палиндром, поскольку симметрия была заложена в алгоритм определения мотива по

данным связывания ER. Из рисунка видно, что мотив состоит из двух половин с пропуском нескольких позиций между ними, что соответствует представлению о связывании белка в форме димера.

Используя такой уточненный палиндромный мотив, определенный по свободной энергии связывания, все 16,043 участков связывания ER были исследованы на присутствие различных групп (субпопуляций) сайтов связывания, таких как полные сайты, полу-сайты и «не-сайты». Нет заранее определенного способа разделения полных палиндромных сайтов и полу-сайтов на основе только оценки сходства («скора») 17-мерного палиндромного мотива. Поэтому G-скор 17-мера был представлен в виде двух компонент - скоров левой и правой половины сайтов GL и GR. В этой схеме подсчета скора, исходные палиндромные сайты связывания должны будут иметь высокую аффинность для левой и правой половины, т.е. оба скора GL и GR будут низкими (близкими к нулю). С другой стороны полу-сайт связывания ERE будет иметь хороший скор только на одной половине 17-мера, и плохой скор на другой половине (например, GL может быть низким, а GR - высоким). С помощью такой декомпозиции можно отделить обогащение скора полу-сайта от обогащения скора полного сайта в нуклеотидных последовательностях ChIP-seq пиков ER, и таким образом, оценить относительный вклад полу-сайтов и полных сайтов в геномное связывание ER.

Для анализа представленности мотива сайтов связывания ER, определенных в эксперименте ChIP-seq в двумерном пространстве левой и правой половины ERE, было рассчитано распределение скоров для всех исследуемых сайтов. Показано что распределение мотивов геномных сайтов связывания ER симметрично относительно левой и правой половины мотива – то есть ярко выражена только половина одна сайта или весь палиндромный мотив целиком. Часть сайтов, определенных в ChIP-seq эксперименте не имела выраженного мотива связывания ни для целого мотива ERE, ни для полусайтов.

Таблица представляет Позиционно-специфичную матрицу энергии связывания PSEM (Position Specific Energy Matrix). Фактор шкалирования τ = 2.39E-07. 17-мер содержит левую (L) и правую (R) половины сайта ERE.

Таблица 3.6

Позиционно-специфичная матрица энергии связывания (PSEM) для мотива ERE

Позиция в мотиве	Позиция полусайта	A	T	G	C
1	L1	0.6791	0.6132	0.8886	0
2	L2	0	1.5384	0.2713	0.8043
3	L3	1.3348	1.2546	0	3.246
4	L4	1.4382	1.6328	0	3.439
5	L5	1.2157	0	0.6739	1.9512
6	L6	2.9798	1.4546	1.8694	0
7	L7	0	1.8863	0.7029	2.2217
8	0	0	0	0	
9	0	0	0	0	
10	0	0	0	0	
11	R1	1.8863	0	2.2217	0.7029
12	R2	1.4546	2.9798	0	1.8694
13	R3	0	1.2157	1.9512	0.6739
14	R4	1.6328	1.4382	3.439	0
15	R5	1.2546	1.3348	3.246	0
16	R6	1.5384	0	0.8043	0.2713
17	R7	0.6132	0.6791	0	0.8886

На основе такого распределения «скора» были выделены состояния полного мотива («Full-ERE»), мотива полусайта («half-ERE»), и отсутствия мотива («No-ERE»). Дополнительно были определены промежуточные категория между полным мотивом ERE и половиной сайта («intermediate full ERE»), когда есть симметрия в скорах между левой и правой половинами, но общий «скор» недостаточен для подтверждения присутствия полного мотива ERE. Наконец те сайты, которые не попали ни в одну из представленных категорий, но показывали умеренный «скор» для левой или правой половины сайта, были классифицированы как промежуточные полусайты («intermediate half ERE»).

Для исследования транскрипционных факторов, которые могут модулировать связывание ER в геноме MCF-7, с помощью программы MDscan был выполнен поиск

нуклеотидных мотивов в геномных последовательностях размером 100 нт, центрированных на ChIP-seq пиках ER. Для идентификации мотивов помимо палиндромного мотива ERE и его вариантов полусайтов, все выделенные ранее полусайты были маскированы (удалены из анализа) в последовательностях, обрабатываемых программой MDscan. Этот поиск выявил мотивы транскрипционных факторов CACD (близкий мотив к SP1), AP1, Forkhead и AP2 как потенциальные ко-мотивы в участках связывания полусайтов ERE и участков отсутствия мотива ERE (“no-ERE”). Для независимой оценки обогащенности этих ко-мотивов в различных группах последовательностей, связываемых ER, для определения связывания перечисленных выше транскрипционных факторов, выявленных с помощью MDscan, использовались весовые матрицы TRANSFAC (символ “V$” в обозначении идентификаторов позиционных весовых матриц позвоночных, не указан): CACD/SP1 представлен матрицей CACD_01, AP1 представлен AP1_C, Forkhead представлен HNF3ALPHA_Q6 и AP2 представлен AP2ALPHA_03. Порог сходства для этих четырех моделей мотивов TRANSFAC был установлен так, чтобы максимизировать представленность мотивов в 16,043 сайтах связывания ER (100 нт район пиков ChIP-seq) относительно случайно выбранных геномных районов того же размера. Используя такое определение представленности мотива был установлен следующий тренд: вместе с уменьшением качества мотива связывания ERE от полных сайтов к промежуточным полу-сайтам, встречаемость ко-мотива в участках связывания ER увеличивалась для всех четырех мотивов. Обратное отношение между качеством мотива ER и встречаемостью ко-мотива находится в соответствии с моделью, когда сайты ERE с низкой аффинностью более вероятно нуждаются в других транскрипционных факторах для связывания ER. Такое содействие может быть в форме прямых белок-белковых взаимодействий, как было предложено для AP-1 и ER [344], или непрямой кооперации или роли в модификации хроматина, как было предложено для FOXA1 [320].

Доли мотивов различных категорий для сайтов, специфичных в клеточных линиях MCF-7 и T47D (т.е. обнаруженных с помощью ChIP-seq в одной линии, но не в

другой) распределилсть так: 3335 сайтов связывания ER являются общими для обеих клеточных линий, 12707 сайтов специфичны для MCF-7- и 1685 сайтов специфичны для T47D. Отметим, что фракция полных мотивов ERE значительно выше в сайтах, общих для двух линий. Для клеточной линии MCF-7 фракция полных сайтов также выше, чем для T47D специфичных сайтов связывания ERα, найденных с помощью ChIP-seq.

Из рисунка 3.11 видно соответствие мотива связывания узнаваемому консенсусу GGTCA для полных сайтов и полусайтов и вырождение мотива для сайтов без мотива.

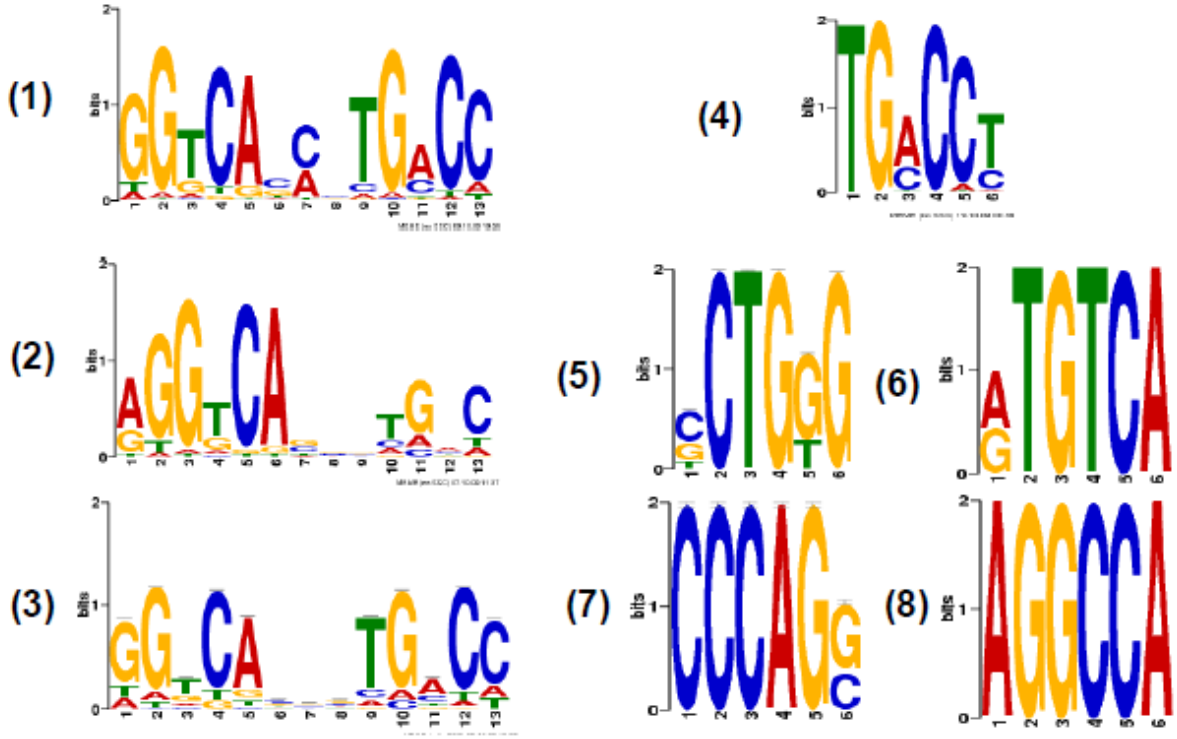


Рис. 3.11. Мотивы связывания ER по категориям (полные сайты, промежуточные полные сайты, полу-сайты и промежуточные полу-сайты, а также сайты без мотива «no ERE»).
1. Лучший мотив полного сайта, найденный программой MEME. 2. Лучший мотив промежуточного полного сайта. 3. Лучший мотив, найденный MEME для промежуточного полного сайта при ограничении построения мотива только для палиндромов. 4. Лучший мотив, найденный MEME для полусайта. 5. Лучший мотив, найденный MEME для промежуточного полусайта. 6. Лучший мотив, найденный MEME для промежуточного полусайта при инициализации поиска с консенсусной последовательности AGGTCA. 7. Лучший мотив, найденный MEME для сайтов без ERE мотива («no-ERE motif»). 8. Лучший мотив, найденный MEME для сайтов без ERE мотива при инициализации поиска с консенсусной последовательности AGGTCA (не показывает узнаваемого мотива ERE).

Мотивы были определены de novo, используя программу MEME, в последовательностях 100 нт вокруг пиков связывания ER ChIP-seq в пяти заданных категориях сайтов. Для полных и промежуточных полных сайтов поиск MEME был ограничен мотивами 13 нт; для других категорий поиск был ограничен размером 6 нт.

Заключение к разделу.

По данным ChIP-seq построена карта сайтов связывания ER в геноме человека. Построен набор потенциальных геном мишеней этого транскрипционного фактора в геноме человека.

В целом, построены геномные карты связывания ТФ MYC, ER, FOXA1 и выполнена компьютерная интеграция данных о сайтах связывания этих ТФ с микрочиповыми данными экспрессии генов в геноме человека [9, 13].

Использование рецептора эстрогенов альфа ERα в качестве модельной системы для детального исследования и определения параметров, определяющих связывание транскрипционного фактора с последовательностями ДНК, позволило уточнить палиндромный мотив связывания по данным ChIP-seq. Показана роль ТФ FoxA1 для активации сайтов связывания ERα.

3.4. Распределение сайтов связывания транскрипционных факторов плюрипотентности по данным ChIP-seq

Использование разработанных компьютерных программ определения положения сайтов связывания транскрипционных факторов в геноме по данным ChIP-seq выполнено автором в серии работ для ТФ плюрипотентности в эмбриональных стволовых клетках.

Понимание строения регуляторных контуров транскрипции, действующих в ЭСК, является фундаментальной основой понимания молекулярной природы плюрипотентности, самообновления и репрограммирования клеток. Несмотря на критичную роль транскрипционных регуляторов в поддержании ЭСК, до сих пор отсутствовала детальная информация об их генах-мишенях in vivo. Мишени вторичных эффекторов ключевых путей передачи сигнала недостаточно изучены, и гены-мишени многих ТФ в ЭСК не были определены.

Использовалась технология ChIP-seq для картирования in vivo участков связывания 13 специфичных к ДНК транскрипционных факторов и других транскрипционных корегуляторов в живущих ЭСК мыши. Оказалось, что эти ТФ связаны в сети взаимодействий в двух основных компонентах. Первая группа факторов включает Nanog, Oct4, Sox2, Smad1, и STAT3. Вторая группа состоит из c-Myc, n-Myc, Zfx, и E2f1. Коактиватор p300 в основном рекрутируется в тесные локусы связывания вместе с белками, найденными в первой группе.

Анализ показал, что эти тесно заполненные локусы связывания имеют характерные черты энхансеосом. ЭСК-специфичная экспрессия генов ассоциирована со связыванием многих изученных факторов. На основе этих ассоциаций между связыванием и экспрессией, было сконструирована модель регуляторной генной сети, которая интегрирует два ключевых сигнальных пути и с внутренними факторами в ЭСК.

Выбор транскрипционных факторов для определения профилей их связывания в ЭСК мыши с помощью ChIP-seq был основан на следующем. Nanog, Oct4, Sox2, Esrrb и Zfx - известные регуляторы плюрипотентности, а также самообновления клеток. Smad1 и STAT3 - ключевые компоненты сигнальных путей, опосредованных BMP и LIF, соответственно. Транскрипционный фактор Tcfcp2l1 имеет повышенную регулируемую экспрессию в ЭСК, но его функция и свойства ДНК-связывания оставались не охарактеризованными [440]. Фактор E2F1 хорошо известен из-за его роли в регуляции прогрессии клеточного цикла, показана ассоциация участков его связывания с промоторными районами генов [444]. ТФ Klf4 и Myc - это факторы репрограммирования, входящие в четверку факторов репрограммирования так называемого «коктейля Яманака». Эти факторы также вовлечены в поддержание недифференцированного состояния ЭСК [436, 445]. Инсулятор CTCF необходим для ограничения транскрипции генов в изолированных областях генома [407]. Картирование сайтов связывания этих 13 ТФ было необходимо для исследования полногеномного связывания факторов в ЭСК, формирования представления об их взаимодействиях между собой.

Иммунопреципитация хроматина со специфичными антителами к этим ТФ с последующим высокопроизводительным секвенированием выполнялась на платформе Illumina (Genome Analyzer). Используя методы и компьютерные программы, описанные в предыдущей Главе, были построены профили связывания, определены пики профилей (кластеры пересекающихся по расположению в геноме ChIP фрагментов

ДНК), статистически значимые пики рассматривались как предполагаемые сайты связывания.

Было идентифицировано от 1,126 до 39,609 сайтов связывания транскрипционных факторов (ССТФ) для этих 13 факторов. В качестве примера, на рисунке показаны профили связывания для всех 13 в локусах, содержащих ген *Pou5f1* (Oct4) (левая панель) и ген *Nanog* (правая панель).

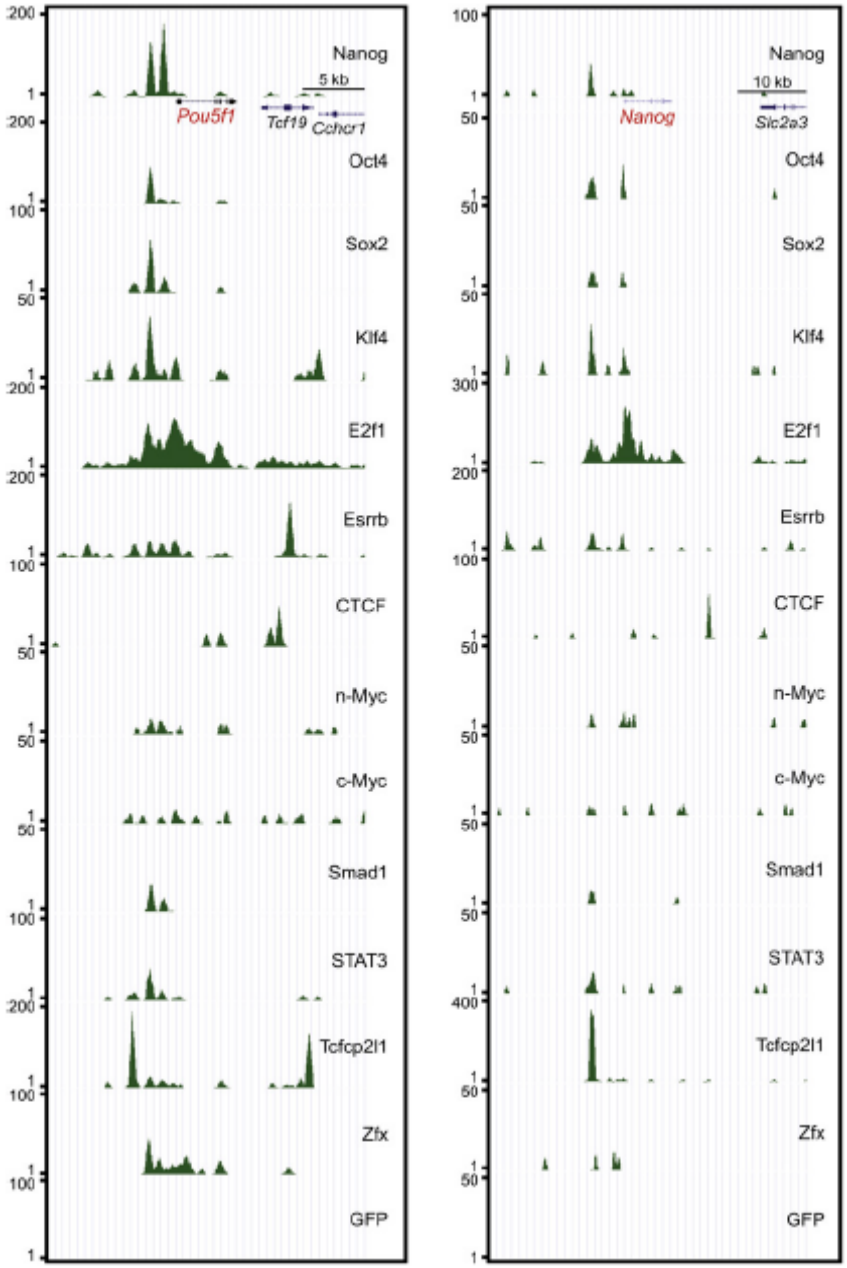


Рис. 3.12. Профили связывания 13 различных транскрипционных факторов в геноме мыши в хромосомных координатах генов Pou5f1 (левая панель) и Nanog (правая панель). Внизу представлен профиль контрольного секвенирования (для неспецифического белка GFP) [3].

Уровни насыщения экспериментальных данных (полноты эксперимента) по определению всех сайтов в геноме были оценены с помощью компьютерной симуляции распределения сайтов в геноме (процедура Монте-Карло), также детально описанной в предыдущей Главе. Результаты моделирования показали достаточную полноту экспериментов по определению сайтов в геноме для всех транскрипционных факторов.

Дополнительно было выполнено ChIP-qPCR тестирования связывания для нуклеотидных последовательностей выборки сайтов с различной интенсивностью связывания (высотой пика), которое подтвердило корректность выбранных пороговых значений ChIP-seq для определения сайтов связывания [3]. Это тестирование показало, что специфичность связывания составила более 95% для большинства ChIP-seq библиотек.

Контекстная структура сайтов связывания транскрипционных факторов (ССТФ)

Вопрос о контекстном составе ССТФ может быть рассмотрен с точки зрения структурных отличий между сайтами и их локальным окружением. На ЭСК мыши были проведены эксперименты хроматин-иммунопреципитации с последующим секвенированием (ChIP-seq) для ТФ Nanog, Oct4, Sox2, Klf4, E2f1, Esrrb, CTCF, n-Myc, c-Myc, Smad1, STAT3, Tcfcf21l, Zfx, Suz12 и контрольное секвенирование (с неспецифичным к ДНК белком GFP) [3].

Для исследования специфичности нуклеотидных последовательностей, связанных с этими ТФ in vivo, были определены мотивы связывания с использованием алгоритма определения мотива de novo (без исходных предположений), описанного ранее в (Loh et al., 2006). Для каждого ТФ был построен список геномных позиций, ранжированный по высоте пика ChIP-seq в эксперименте соответствующем связыванию исследуемого фактора. Для каждого фактора из наибольших по высоте 500 пиков были взяты нуклеотидные последовательности (±100 нт) от центра пика. Повторы в этих последовательностях были маскированы (помечены на присутствие повторяющихся геномных последовательностей программой RepeatMasker), и не использовались в дальнейшем анализе. Использовалась программа Weeder (Pavesi et al.,

2001) для поиска сверхпредставленных последовательностей олигонуклеотидов, из которых были построены мотивы (весовые матрицы). Высокое разрешение профилей связывания позволило выделить сверхпредставленные мотивы для 12 из 13 факторов (исключая E2f1). В соответствии с более ранней работой [429], был получен композитный элемент *sox-oct,* состоящий из консенсуса сайта связывания Sox2 (5′-CATTGTT-3′) и канонической последовательности связывания Oct4 (5-′ATGCAAAT-3′) присоединенных друг к другу в обоих наборах данных Oct4 и Sox2. Присутствие общего мотива дает возможность предположить, что гетеродимер Sox2 и Oct4 является функциональной единицей связывания.

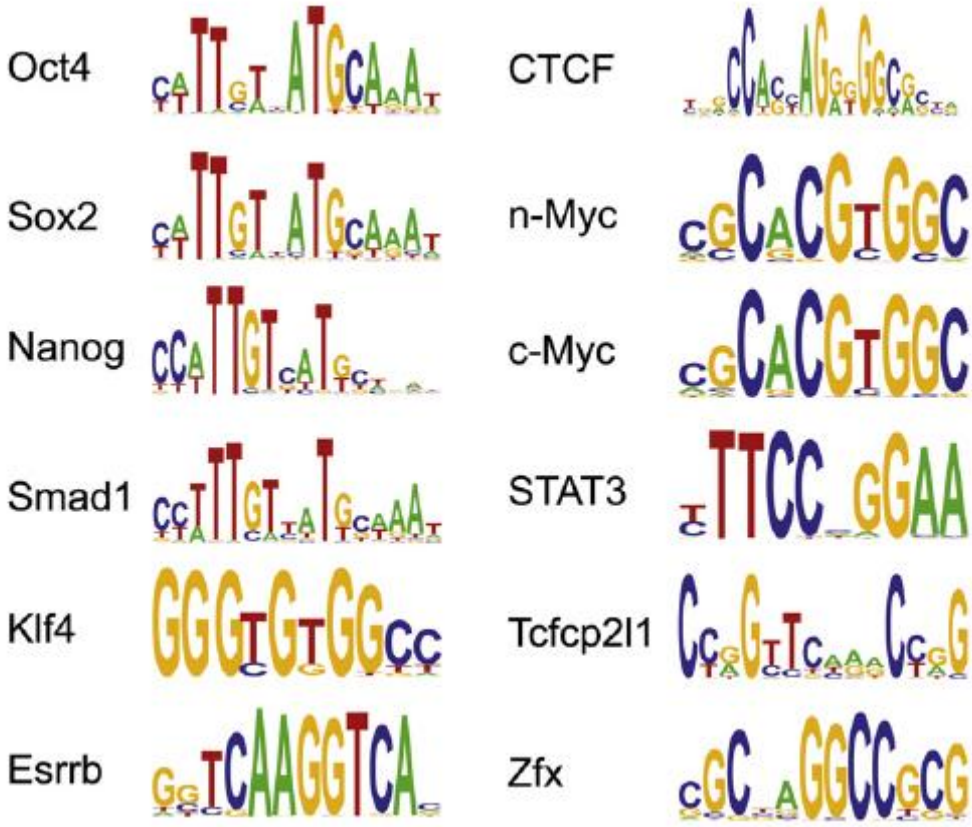


Рис. 3.13. Уточненные мотивы ССТФ определенные по полногеномным данным ChIP-seq для ЭСК мыши [3].

Интересно отметить, что предсказанные de novo матрицы связывания для Nanog и Smad1 повторяют совместный мотив *sox-oct*. Это отражает частоту совместного связывания Nanog и Smad1 с Sox2 и Oct4. Следует отметить, что мотив Nanog, установленный ранее [429] может быть найден с помощью другого алгоритма

определения мотива - NestedMICA. Консенсусные последовательности связывания, найденные для транскрипционных факторов Klf4, Esrrb, CTCF, c-Myc, n-Myc, STAT3 и Zfx имеют близкое сходство с последовательностями, представленными ранее [9, 318, 407, 435, 441]. Таким образом, нуклеотидные мотивы связывания могут быть определены из полногеномных данных связывания in vivo, что принципиально улучшает возможности исследования регуляторных районов транскрипции.

Отметим, что мотивы связывания для факторов c-Myc и n-Myc мыши близки между собой и соответствуют мотиву связывания MYC человека [9], описанному ранее в данной Главе.

3.5 Регуляторные контуры взаимодействий генной сети по данным связывания транскрипционных факторов

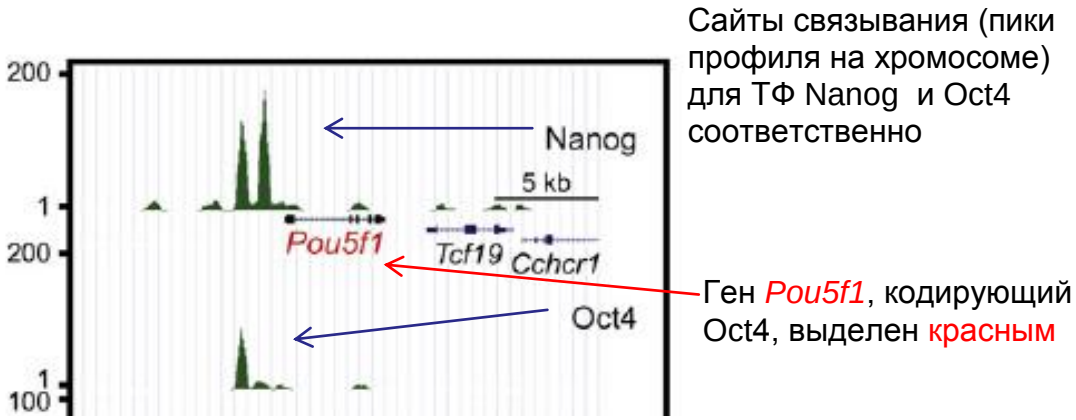
Показано, что ключевые факторы плюрипотентности Oct4, Nanog, Sox2 образуют тесно связанную регуляторную сеть. Возможна реконструкция регуляторных контуров ТФ на основе расположения сайтов связывания в промоторах генов-мишеней и данных экспрессии генов на микрочипах.

Статус самообновления недифференцированных ЭСК характеризуется экспрессией генов, специфически регулируемых в этом типе клеток. Задача определения регуляторной генной сети, определяющей ЭСК-специфичную экспрессию, может быть решена через использование сайтов связывания транскрипционных факторов, которые связываются в регуляторных районах своих генов мишеней, также являющихся транскрипционными факторами.

Для выявления регуляторных взаимодействий, связанных с экспрессией генов, использовались опубликованные данные экспрессии на микрочипах, содержащие сравнение недифференцированных клеток с клетками, начавшими процесс дифференцировки. Известно, что 9 из исследованных 13 факторов, (а именно Nanog, Oct4, Sox2, Klf4, n-Myc, c-Myc, Esrrb, Zfx, Tcfcp2l1) снижают уровень экспрессии при дифференцировке или в других, уже специализированных типах клеток [440]. Использовались два набора микрочиповых данных для определения генов, дифференциально экспрессирующихся (т.е. имеющих значительную разницу в уровнях экспрессии) при дифференцировке клеток [440, 450]. Сравнение двух независимо

полученных наборов данных минимизирует возможные ошибки в определении дифференциальной экспрессии генов из-за различий в способах дифференцировки клеток.

Для определения регуляторной сети между исследуемыми ТФ и их генами мишенями был составлен список генов, вовлеченных в такую сеть, используя списки связанных генов (имеющих сайт связывания - пик ChIP-seq - в окрестностях гена) и списки дифференциально экспрессирующихся генов по данным микрочипов. Список генов-мишеней действия ТФ в ЭСК мыши был ранжирован по силе связывания (общему числу высот пиков ChIP-seq в окрестностях гена). Был ранжирован и список дифференциально экспрессирующихся генов по статистической значимости разницы в уровнях экспрессии между плюрипотентным состоянием клеток, и клеток, начавших дифференцировку. Порог для определения верхних генов из объединенного ранжированного списка (пересечения двух списков) был выбран следующим образом: должно было быть как минимум в два раза больше генов в пересечении, чем ожидается по случайным причинам (по сравнению с нулевой гипотезой, когда пересечение списков случайно, такая гипотеза должна быть отвергнута на уровне значимости $p<10^{-3}$). Такой подход позволяет использовать все данные и избежать использования единственного порогового значения для всех библиотек ChIP-seq.



Элемент регуляторной сети

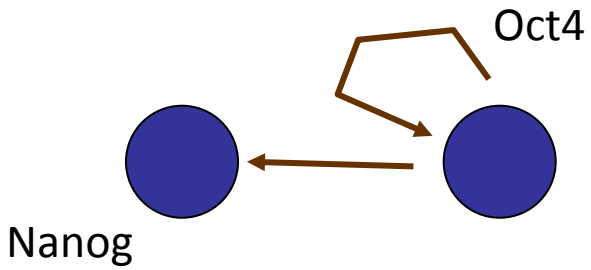


Рис. 3.14. Профили связывания факторов Nanog и Oct4 в районе гена Pou5f1 (кодирующего сам Oct4) по данным ChIP-seq в ЭСК мыши (верхняя панель). Реконструкция регуляторных воздействий в генной сети транскрипционных взаимодействий для представленных генов Oct4 и Nanog: саморегуляция для Oct4 (обратная связь на Pou5f1) и направленная регуляция для Nanog (нижняя панель).

Реконструкция регуляторных взаимодействий для нескольких ТФ в геноме мыши с учетом данных экспрессии генов представлена на следующем рисунке по данным [3].

Сеть, построенная из 13 транскрипционных факторов, показанная на рисунке 3.11, выявила как ожидаемые, так и неожиданные аспекты взаимоотношений между этими ТФ. В соответствии с более ранними исследованиями, эта модель показа наличие регуляторных обратных связей между Oct4, Sox2, и Nanog (Boyer et al., 2005, Chew et al., 2005; Loh et al., 2006). Новая черта этой сети - взаимозависимости между большинством из 13 исследованных транскрипционных факторов.

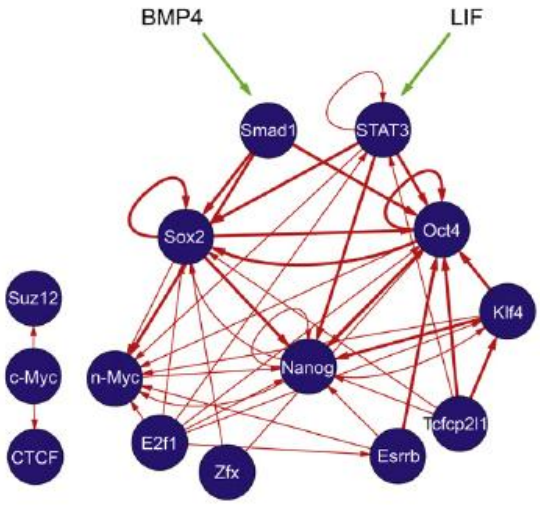


Рис. 3.15. Генная сеть регуляции факторов плюрипотентности и ТФ в эмбриональных стволовых клетках мыши (Chen et al., 2008).

3.6 Энхансеры и множественные локусы регуляции транскрипции по данным ChIP-seq

Используя данные ChIP-seq для профилей связывания ТФ в ЭСК мыши далее исследовались взаимодействия транскрипционных факторов в плане одновременного связывания различных ТФ в геномных районах, размером до 500 нт (т.н. множественные локусы регуляции транскрипции, или MTL (multiple transcription loci) (Chen et al., 2008) (рис. 3.18).

Некоторые районы генома заняты несколькими различными транскрипционными факторами одновременно, связанными с геномной ДНК на очень близком расстоянии (десятки нуклеотидов), или даже перекрываясь своими сайтами связывания. Некоторые обогащенные связыванием ТФ районы могут появиться по случайным причинам - близкое расположение сайтов еще не означает их функциональной общности или кооперативного связывания.

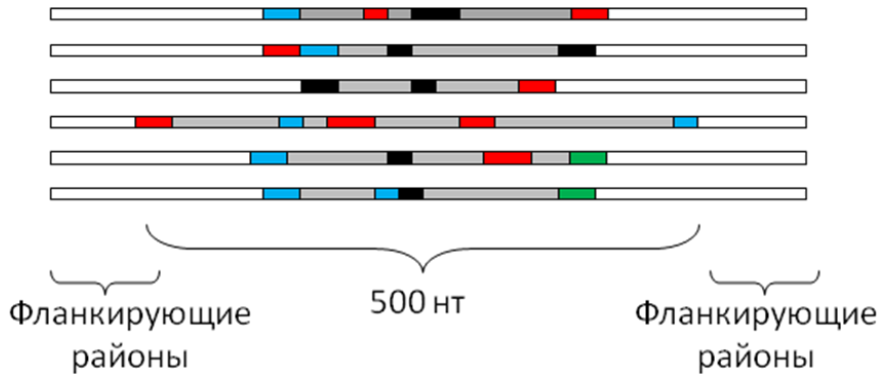


Рис. 3.16. Схематическое представление множественных локусов регуляции транскрипции, или MTL [3].

В то же время некоторые геномные районы, обогащенные сайтами связывания, (Chen et al., 2008), могут функционировать как дистальные энхансеры, и действительно привлекают кооперативно связывающиеся белковые факторы, физически контактирующие друг с другом при связывании с ДНК. Для статистического разделения неслучайных комбинаций сайтов от «шума» - ожидаемого по случайным причинам числа кластеров сайтов (пар и групп близко расположенных позиций пиков связывания ChIP-seq), был разработан алгоритм, принимающий во внимание число связанных районов, интенсивность сигнала ChIP-seq в связывании для каждого ТФ. Первый шаг состоял в формальном определении кластера сайтов связывания. Два участка связывания (пики ChIP-seq) включались в кластер, если центральные позиции пиков были удалены не более чем на 100 нт друг от друга.

В целом, 3 или 4 различных сайта связывания ТФ в одном и том же геномном локусе могут рассматриваться как неслучайная комбинация. Для оценки вероятности получения таких комбинаций было построено распределение кластеров, которые могут образоваться по случайным причинам, по размерам, учитывая размеры сайтов и размеры хромосом. Для более строго сравнения использовался только размер генома, занимаемый промоторами, без пересечения (используя определение 2.5Кб перед стартом транскрипции, и 500 нт после), что значительно меньше размера всего генома, доступного для картирования. Связывание 4 и более ТФ одновременно достаточно для принятия гипотезы на уровне 1% FDR (с 1% вероятность ошибки ложного

предсказания), как для проксимальных промоторов, так и для дистальных сайтов. В целом для всей совокупности ССТФ было получено 33337 (~20%) сайтов связывания в промоторных районах и 129647 (~80%) сайтов внутри генов и в дистальных районах. Общий размер без пересечений всех промоторных районов <-2500;+500> нуклеотидов относительно стартов транскрипции для 20,993 генов RefSeq мыши составляет 56.4Мб (точнее 56,428,675), что из-за перекрывания генов в геномных координатах чуть меньше общего размера произведения длин на число промоторов оцениваемого приблизительно в 63Мб (3000*20,993). Отдельная компьютерная симуляция образования случайных кластеров в промоторах, используя размеры хромосом и промоторов, показала, что кластеры размера 4 уже не случайны, с уровнем значимости $p<0.01$. Заметим, что для кластеров сайтов размера 3 в дистальных районах уровень ошибки FDR составляет 1.59%, что является приемлемым, но для промоторных районов ошибка FDR составляет уже 5.60%. В целом, для анализа кластеров сайтов в геноме мыши использовался размер 4 и более, что удовлетворяет критериям статистической значимости как для промоторных, так и для непромоторных сайтов.

Локусы множественного связывания транскрипционных факторов могут рассматриваться как энхасеосома - набор энхансерных районов - в эмбриональных стволовых клетках. При рассмотрении профилей связывания этих 13 транскрипционных факторов, было обнаружено, что многие участки генома связаны несколькими сайтами одновременно. Примером являются показанные на рисунке ранее профили ChIP-seq нескольких транскрипционных факторов для районов генома мыши, содержащих ген Pou5f1 и Nanog.

Перед исследованием их биологического значения найденных кластеров была определена значимость обогащения кластерами сайтов таких небольших районов. Пики ChIP-seq для фиксированного ТФ, содержащие рядом, в окрестности 100 нт, пики другого фактора, последовательно кластеризовались друг с другом. Кластер увеличивался до тех пор, пока новые пики уже невозможно было добавить. Ограничением по длине геномного участка был размер 500 нт. Таблица полученных

кластеров, всего 3583 множественных локусов транскрипции (так называемых MTL), максимально до 11 сайтов в кластере, приведена в Приложении. Для каждого локуса получено описание, сколько сайтов разных ТФ, заданных пиками ChIP-seq, он содержит, имена этих сайтов, высота пиков, его геномные координаты. Вероятность получения таких кластеров может быть оценена индивидуально с учетом того, что число пиков для каждого ТФ свое. Использовался приближенный подход оценки вероятности образования таких кластеров по случайным причинам, когда рассматривалась комбинация разных сайтов всех исследуемых ТФ без учета высоты пика. Рассмотрим функцию числа кластеров в геноме в зависимости от числа сайтов в кластере. Такая функция распределения кластеров по размерам показана на рисунке 3.19. Для сравнения приведена функция числа кластеров, которые могут наблюдаться по случайным причинам. Распределение было рассчитано с помощью специально написанной для этого компьютерной программы, учитывающей размер генома мыши и размеры кластера (до 100 нт между пиками). Заметим, что ожидаемое по случайным причинам распределение можно получить с помощью распределения Пуассона, если представить весь геном как линейную последовательность дискретных участков - отрезков (размера до 500 нт), или «бинов» - дискретных единиц и считать вероятность попадания двух сайтов из общего набора в такие отрезки. Если разделить число кластеров, ожидаемое по случайным причинам, к числу, наблюдаемому в эксперименте, получим вероятность ошибки или статистическую значимость кластера данного размера. Показано, что кластеры, содержащие 4 и более сайтов, имеют высокий уровень значимости ($p<0.001$, см. рисунок); всего получено 3583 таких кластера. Кластеры из двух и трех сайтов в геноме уже могут быть получены по случайным причинам ($p>0.05$).

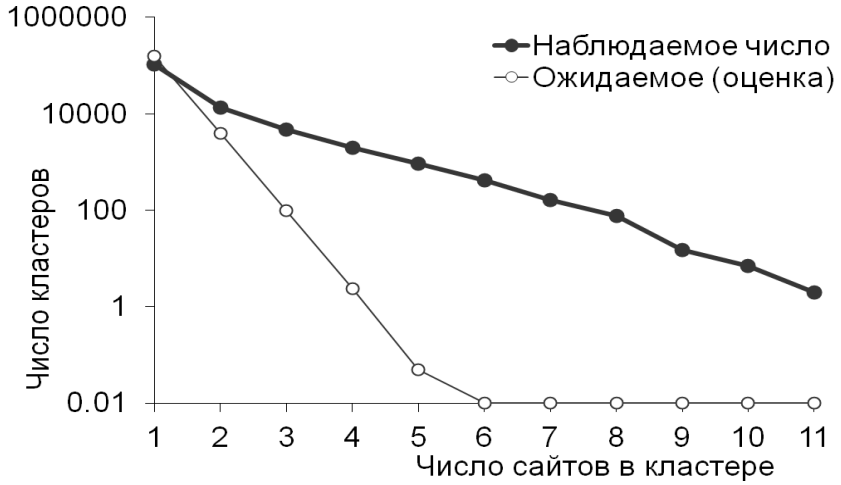


Рис. 3.17. Оценка встречаемости кластеров сайтов связывания ТФ в эмбриональных стволовых клетках мыши [3].

Из отобранных кластеров сайтов 1440 (40.2%) находятся во внутригенных районах, оставшиеся кластеры располагаются в промоторных районах (1334 локуса, 37.2%) и внутри генов (809 локуса, 22.6%). Интересно отметить, что кластеры большего размера, как правило, расположены дистально - лишь менее 20% кластеров из 7 и более ТФ находятся в промоторных районах (см. рисунок), в сравнении с 40% кластеров размера не более 5 ТФ. Следовательно, совместная встречаемость ССТФ в кластерах не связана с их расположением в промоторах.

Показано, что кластеры сайтов, состоящие из 3 и более сайтов, могут рассматриваться как неслучайные (p<0.01). Для промоторных районов неслучайными являются кластеры сайтов различных ТФ от 4 и более сайтов.

Из рисунка видно последовательное увеличение доли дистально расположенных кластеров при увеличении числа сайтов в кластере. В то же время среди больших кластеров, содержащих 7, 8 и более сайтов, значительно меньшая доля располагается в промоторных районах генов. Интервал [-2500;+500] относительно старта транскрипции был использован как определение промоторного района, следуя определению, предложенному в работе [369].

Видно, что группа кластеров Myc (включая гены c-Myc, n-Myc). характеризуется

преимущественно промоторным расположением, а группа Nanog (с другими ключевыми факторами плюрипотентности) - преимущественно дистальным расположением относительно старта транскрипции генов.

Для каждой пары ТФ в геноме был рассчитан коэффициент корреляции Пирсона совместной локализации их сайтов связывания в геноме мыши. Была построена матрица частот совместной локализации сайтов различных ТФ. Для каждой пары транскрипционных факторов с помощью компьютерной программы рассчитывалось число перекрывания в геноме пиков профилей ChIP-seq. В каждой ячейке такой симметричной матрицы получается число совпадающих (перекрывающихся в геномном интервале до 200 нт) сайтов связывания двух факторов. По диагонали располагается число сайтов. Матрица была переведена в матрицу корреляций. Для каждой пары транскрипционных факторов был подсчитан линейный коэффициент корреляции Пирсона – корреляции между строками исходной матрицы, соответствующих числам пересечений геномных позиций с сайтами других факторов. Такая матрица задает меру близости, или ассоциации, между расположением сайтов связывания различных транскрипционных факторов. Чем выше коэффициент корреляции, тем более близко расположение сайтов связывания двух транскрипционных факторов между собой относительно других факторов. Может использоваться также ранговый коэффициент корреляции. Используя эту меру с помощью программ среды R была выполнена кластеризация исследованных транскрипционных факторов друг с другом.

Был проанализирован состав транскрипционных факторов в 3583 кластерах. В целом в исследованном наборе выделяются две группы: относящиеся к Myc и относящиеся к Nanog транскрипционные факторы.

Среди 13 факторов, Nanog, Sox2, Oct4, Smad1, и STAT3 имеют тенденцию встречаться совместно более часто, также выделяется вторая группа, состоящая из факторов n-Myc, c-Myc, E2f1 и Zfx (рис. 3.22). В дополнение к этим двум основным группы кластеров можно выделить промежуточную группу кластеров, содержащих

сайты связывания из основных групп, то есть на основе присутствия или отсутствия сайтов (1) Oct4, Sox2, или Nanog и (2) c-Myc или n-Myc. Кластеры Nanog-Oct4-Sox2 (где связывание наблюдается для Nanog, Oct4, или Sox2, но не для n-Myc или c-Myc, составляют 43.4% от общего набора 3583 кластеров. Myc-специфичные кластеры (n-Myc или c-Myc, но не Nanog, Oct4, или Sox2) составляют 32.9% всех кластеров MTL (см. рисунок).

В соответствии с попарной встречаемостью, показанной на рисунке, 87.4% сайтов связывания Smad1 и 56.8% сайтов STAT3 в кластерах MTL ассоциированы с Nanog-Oct4-Sox2-специфичными MTL. Такая ассоциация сайтов указывает на то, что Smad1 и STAT3 разделяют многие общие регуляторные районы с ТФ Nanog, Oct4, и Sox2 и отражает точку сходства двух ключевых путей передачи сигнала (через Smad1 и STAT3) с основным регуляторным контуром ЭСК, определяемым тройкой факторов Nanog, Oct4, Sox2 [428]. Такое наблюдение соответствует ранее опубликованной работе, показавшей связь между путями передачи сигнала Nanog и LIF [433]. В целом 56.9% сайтов связывания Esrrb и 41.9% сайтов связывания Klf4 находятся в Nanog-Oct4-Sox2-специфичных кластерах MTL. Действительно, ранее было показано, что ТФ Esrrb находится в том же белковом комплексе, что и Nanog [442]. Напротив, совместная встречаемость сайтов связывания Zfx, CTCF и E2f1 смещена в сторону Myc-специфичных.

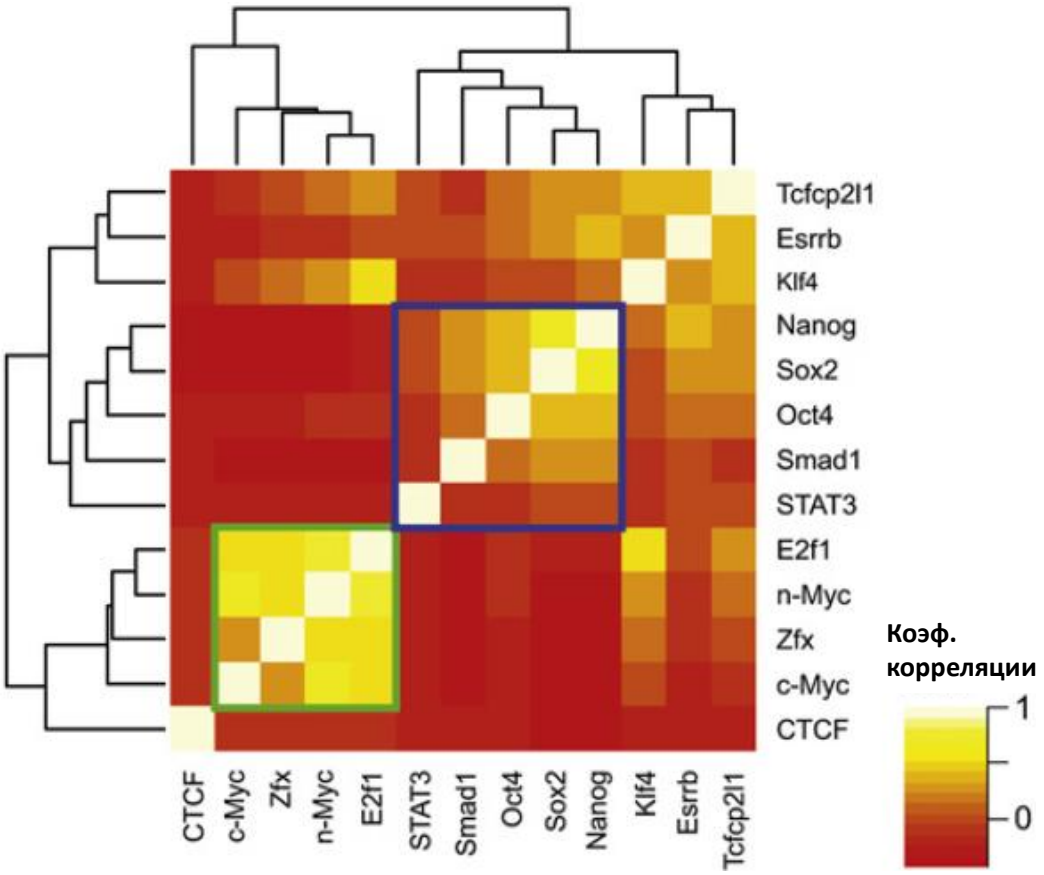


Рис. 3.18. Термокарта кластеризации ССТФ в ЭСК мыши для 13 факторов [3].

Поскольку большинство Nanog-Oct4-Sox2-специфичных кластеров найдены вне промоторных районов (91.2%), было выполнено тестирование геномных последовательностей из этих типов кластеров на энхансерную активность. Всего 25 геномных фрагментов из кластеров Nanog-Oct4-Sox2 и 8 геномных фрагментов из кластеров Myc были клонированы в репортерной конструкции под промотором люциферазы. Геномные фрагменты были помещены в 2Кб от минимального промотора *Pou5f1* используемого для экспрессии гена *luciferase*. Эти конструкции были внедрены в ЭСК и клетки линии 293T, затем измерялась люциферазная активность. Интересно отметить, что все 25 конструктов с геномными фрагментами, содержащими кластеры Nanog-Oct4-Sox2, показали устойчивую энхансерную активность, специфичную для ЭСК. В противоположность этому, контрольные конструкции с геномными фрагментами из кластеров Myc были неактивны или показывали очень слабую ЭСК-специфичную активность [3].

Комбинации различных транскрипционных факторов, связывающихся с энхансером, могут синергично влиять на транскрипцию и на связывание друг друга [298]. Для исследования взаимоотношений между Oct4, Smad1 и STAT3, дополнительно был выполнен эксперимент по нарушению связывания этих факторов через интерференцию РНК (RNAi) и прекращению действия ростовых факторов. Подавление Oct4 приводило к уменьшению связывания Smad1 и STAT3. Перемена связывания Smad1 и STAT3 проявлялась специфично на совместно связанных сайтах Oct4, Smad1 и STAT3, но не из-за уменьшения уровней связывания Smad1 и STAT3. Были выполнены обратные эксперименты по прекращению действия LIF и BMP4 из среды культивирования. Вывод из среды фактора LIF уменьшал связывание STAT3 со своими мишенями, тогда как вывод BMP4 уменьшал связывание Smad1 с его мишенями. Экспериментальное подавление этих двух сигнальных путей, тем не менее, не имело эффекта на связывание Oct4 (ChIP эксперимент на шести тестовых последовательностях по связыванию Oct4 с антителом на экстрактах ЭСК, обработанных соответственно только LIF, только BMP4, LIF совместно с BMP4 или в их отсутствие). Эти результаты указывают на то, что Oct4 имеет ведущую роль в стабилизации ДНК-белкового комплекса и устанавливает иерархию регуляторных взаимодействий между Oct4, STAT3 и Smad1. Механизм Oct4-зависимого связывания STAT3 и Smad1 в настоящее время не ясен. Возможно, что Oct4 может взаимодействовать с STAT3 и Smad1 для облегчения их связывания с хроматином.

В целом, через глобальный анализ профилей связывания ТФ было определено свыше 3000 геномных районов, с тесным расположением сайтов связывания различных транскрипционных факторов. Кластер сайтов Nanog-Oct4-Sox2 имеет характерные черты энхансеосомы поскольку усиливает сигнал транскрипции с большого расстояния от генов, а также показывает широкое совместное связывание с Smad1 и STAT3. Важно отметить, что Oct4 необходим для связывания Smad1 и STAT3, что подтверждает, что Oct4 играет основную роль в стабилизации комплекса транскрипционных факторов.

Дополнительный анализ кластеров MTL: связывание p300 и Suz12. В дополнение к ChIP-seq экспериментам по определению полногеномного распределения сайтов связывания 13 транскрипционных факторов, в ЭСК мыши с помощью ChIP-seq были определены сайты связывания транскрипционного коактиватора p300. p300 - это гистон ацетилтрансфераза, часто присутствующая в энхансерных районах [369]. Полногеномное картирование таких регуляторов хроматина как p300 помогает выявить ДНК-связывающие факторы, ответственные за рекрутирование (привлечение) регулятора к специфическим сайтам генома [370]. В качестве контроля был определен профиль связывания другого регулятора хроматина, Suz12.

Было установлено, что сайты связывания p300 встречаются в кластерах группы Nanog-Oct4-Sox2 (84.2% сайтов p300 находится в кластерах этой группы). Большинство сайтов связывания p300 ассоциированы с 3–6 другими ТФ, в максимальном варианте до 9 ТФ. Композиция кластеров, содержащих p300 очень разная, типично они включают один или более сайтов связывания факторов Nanog, Oct4 или Sox2, включая затем, с меньшей вероятностью Smad1, Esrrb, Klf4, Tcfcp2l1 и STAT3. В противоположность p300, сайты Suz12 не показали достаточно строгой ассоциации с каким-либо из 13 первоначально рассмотренных ТФ (данные не показаны).

На рисунке 3.19 показаны профили связывания ChIP-seq, включая p300 и Suz12 в районе гена *Loxl1* и гена *Lefty1* в геноме мыши.

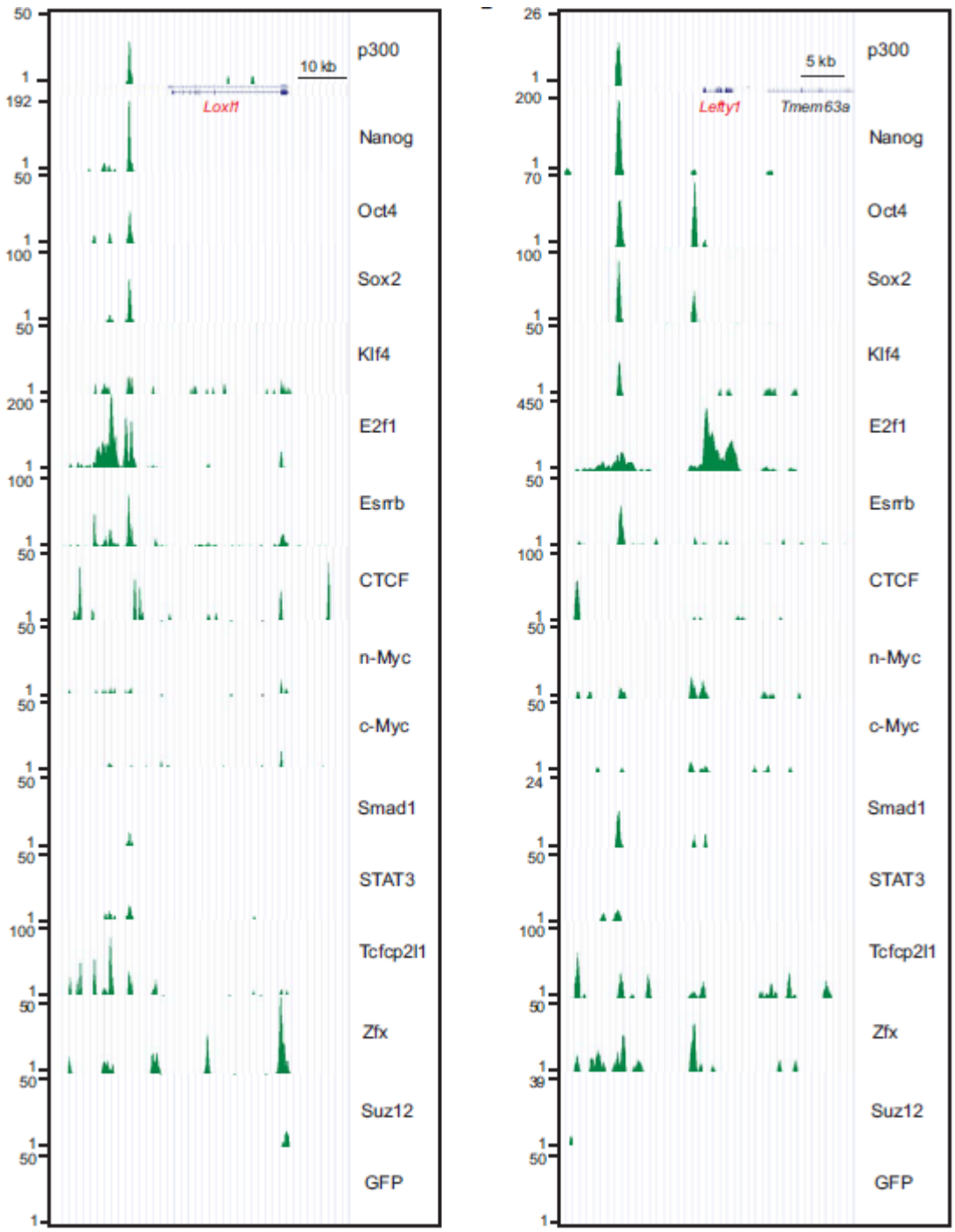


Рис. 3.19. Пример расположения ChIP профилей связывания в ЭСК мыши в районе гена Loxl1 (левая панель) и гена Lefty1 (правая панель) для 15 факторов, включая p300 и Suz12 [3].

Используя алгоритм реконструкции de novo мотивов Weeder, по нуклеотидным последовательностям пиков ChIP-seq был определен мотив связывания p300, который близок к элементу связывания *sox-oct* (рис. 3.20).

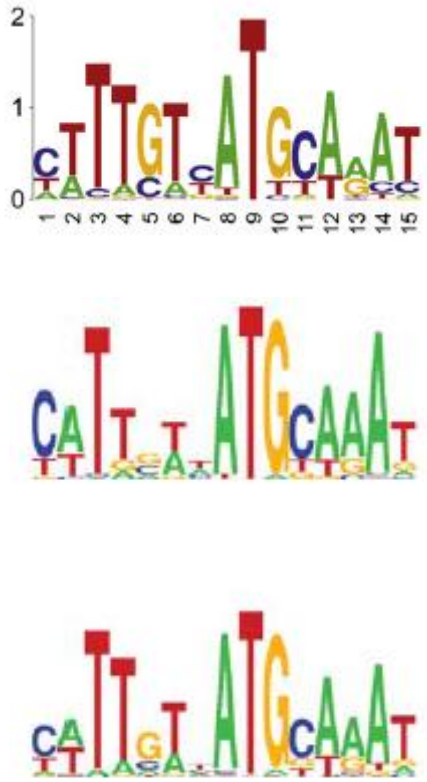


Рис. 3.20. Мотив связывания p300 и мотивы Oct4, Sox2, определенные по данным ChIP-seq в ЭСК мыши.

Ассоциация p300 с кластерами Nanog-Oct4-Sox2 была проверена экспериментально для 12 сайтов с использованием ChIP-qPCR. Такие данные подтверждают, что Oct4, Sox2 и Nanog рекрутируют (привлекают) p300 к своим геномным сайтам связывания. В ЭСК не наблюдалось глобального рекрутирования p300 к сайтам Myc. Подавление экспрессии c-Myc с помощью РНК-интерференции не влияло на рекрутирование p300 к этим сайтам. Эти результаты позволяют предположить, что общий фактор p300 рекрутируется к энхансерам, и это рекрутирование поддерживается Oct4, Sox2 и Nanog [369]. Такое предположение было подтверждено в дальнейшем [371].

3.7 Компьютерное исследование ко-локализации в геноме и построение тепловых карт кластеров сайтов связывания

Для транскрипционных факторов c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2, PRDM14 в геноме мыши и в геноме человека были построены геномные карты.

Показана ко-локализация ССТФ, ответственных за плюрипотентность (Кластер Oct4, включающий факторы из базового набора репрограммирования Oct4-Nanog-Sox2-Klf4). Кластеризация сайтов по взаимному расположению представлена на «тепловой карте» (рис. 3.21).

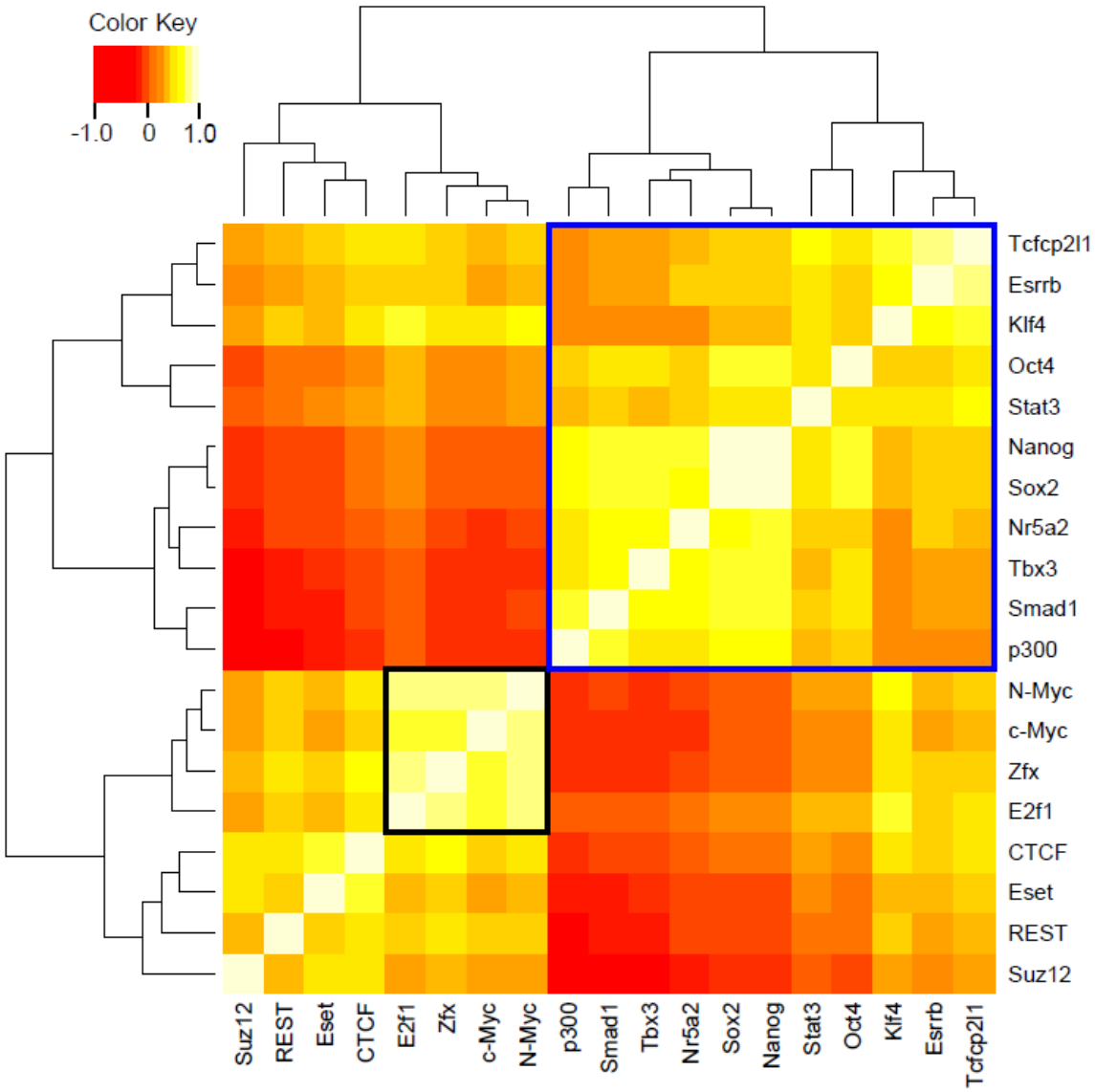


Рис. 3.21. Ко-локализация ССТФ в геноме мыши, включая ТФ Eset, Nr5a2, Tbx3.

Заключение к разделу

Таким образом, выполнено широкомасштабное картирование сайтов связывания большого набора транскрипционных факторов в геноме мыши. Впервые на одном типе клеток в масштабе генома определены сайты связывания 15 транскрипционных факторов одновременно. Отметим, что, несмотря на развитие технологий картирования

для всестороннего и объективного определения всего репертуара сайтов связывания в геноме, трудно установить какие именно из геномных сайтов влияют на транскрипцию. Возможно, что значительная доля этих сайтов нефункциональны и представляют собой следствие биологического шума в передаче регуляторного сигнала [298]. Важно отметить, что эксперименты иммунопреципитации хроматина могут получать сигнал непрямых ДНК-белковых взаимодействий через белок-белковые взаимодействия. Преимущество выполненного анализа состоит в изучении одновременного расположения множества ТФ в клетках одного типа. Результаты показывают, что есть геномные районы, совместно занятые несколькими ТФ (горячие точки совместной локализации ТФ). Такие участки более вероятно представляют собой функционально важные сайты. Отметим более раннее исследование профилей связывания девяти транскрипционных факторов в ЭСК мыши [407]. Выполненное исследование представляет сайты связывания ТФ в масштабе генома используя антитела, распознающие эндогенные белки клетки, тогда как Kim и соавторы [407] использовали промоторные ДНК микрочипы для исследования присутствия белков, связанных биотином. Также проведенное исследование включало анализ расположения в геноме вторичных эффекторов основных сигнальных путей (Smad1, STAT3), регуляторов самообновления (Zfx, Esrrb), инсуляторных белков (CTCF), и транскрипционных корегуляторов (p300, Suz12). Метод ChIP-seq выявил богатый репертуар сайтов связывания в геноме, в том числе расположенных в дистальных районах и вырожденных повторах. Объединение и интеграция данных ChIP-seq позволит найти новые детали функционирования найденных регуляторных районов [37].

Энхансеосома определяется как нуклеопротеиновый комплекс, состоящий из различных наборов сайтов связывания ТФ связанных напрямую или опосредованно с энхансерной ДНК [194]. Плотность сайтов связывания встречающихся на этом коротком сегменте ДНК высока по сравнению с более «модульными» энхансерами, имеющими менее плотно расположенные кластеры сайтов на более длинном сегменте геномной ДНК [195]. Прототипом энхансеосомы может служить вирус-индуцируемый энхансер гена интерферона-β (*IFN-β*). Этот 55 нуклеотидный энхансер связан субъединицами p50 и p65 NF-κB, ATF-2, IRF-3, IRF-7, c-Jun, и архитектурным транскрипционным фактором HMGA. Атомная модель этого комплекса содержащего восемь этих факторов, связанных с ДНК была реконструирована ранее на основе трех кристаллических структур [196]. Анализ этих структур показал ограниченные белок-белковые взаимодействия. Таким образом кооперативность действия

транскрипционных факторов в энхансере может быть опосредована белок-белковыми контактами. Предполагают, что связывание этих восьми ТФ на составной поверхности двойной цепи ДНК создает постоянную площадку для постоянного привлечения коактиваторов, таких как p300 [368].

Проведенный анализ полногеномного картирования ССТФ выявил геномные районы, обладающие характеристиками энхансеосомы. Во-первых, сайты связывания плотно кластеризуются внутри относительно компактных геномных сегментов. В частности наиболее тесно заполненный ССТФ локус был найден в удаленном энхансере гена *Pou5f1*, кодирующего ключевой фактор плюрипотентности Oct4. Этот район [203] связан 11 различными транскрипционными факторами. Во-вторых, было показано, что 25 этих геномных районов функционируют как энхансеры - усиливают транскрипцию при размещении после репортерного гена люциферазы. В-третьих, эти районы ассоциированы с маркером модификации гистонов H3K4me3, который является одним показателей активной транскрипции в геноме. В-четвертых, анализ ChIP-seq для p300 показал глобальное рекрутирование этого коактиватора к кластерам Nanog-Oct4-Sox2, но не к кластерам сайтов Myc. Более того, такое рекрутирование p300 зависит от Oct4, Sox2 и Nanog. У высших эукариот транскрипционные энхансеры играют важную роль в интеграции различных сигнальных путей для активации специфичных генов. Профили множественного связывания ТФ в полногеномной шкале позволили установить совместную локализацию множества различных ССТФ в определенных сайтах в геноме для эмбриональных стволовых клеток.

3.8. Дальнейшие исследования ССТФ в ЭСК мыши с помощью ChIP-seq

Используя геномный анализ генов в ЭСК, играющих роль в плюрипотентности и индуцирующих репрограммирование, было выполнено исследование транскрипционного фактора Tbx3, который существенно улучшает качество ИПСК (индуцированных плюрипотентных стволовых клеток, iPS). Показано, что ИПСК, сгенерированные с помощью набора OSK (Oct4, Sox2 и Klf4, классического набора факторов репрограммирования Яманака) и Tbx3 (назовем такой набор OSKT) имеют лучшие качества как по вкладу клеток зародышевой линии в гонады, так и по частоте передачи зародышевой линии. Тем не менее, глобальный анализ профиля экспрессии не может выявить различия между ИПСК, индуцированными наборами OSK и OSKT. Полногеномный анализ данных иммунопреципитации хроматина с последующим

секвенированием участков связывания Tbx3 в ЭСК мыши показал, что Tbx3 регулирует гены, ассоциированные с состоянием плюрипотентности и факторы репрограммирования. Кроме того, Tbx3 имеет много общих регуляторных мишеней совместно с Oct4, Sox2, Nanog и Smad1.

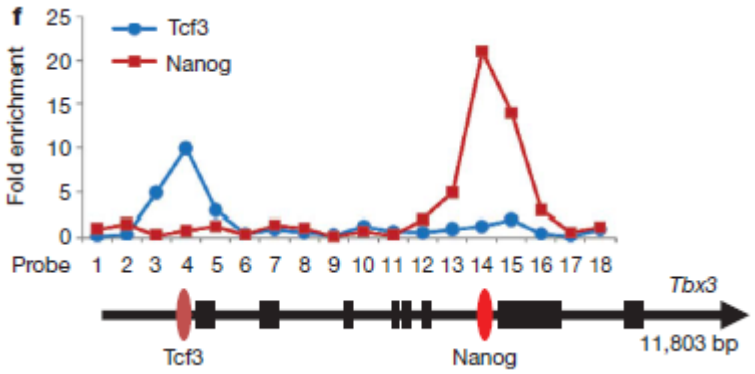


Рис. 3.22. Определение присутствия Tcf3 и Nanog в районе гена Tbx3, измеренное с помощью ChIP-qPCR [41].

Проведенное исследование подчеркнуло качественные различие между ИПСК сгенерированными с помощью различных методов, и показало необходимость тщательной характеризации ИПСК перед изучением in vitro. Свойства плюрипотентности и самообновления ЭСК задаются набором основных факторов, которые помогают определить их уникальную идентичность. Взрослые соматические клетки могут быть репрограммированы чтобы воспроизвести свойства ЭСК когда введены лишь некоторые ключевые транскрипционные факторы. В то же время важна задача повышения эффективности репрограммирования, которая может быть повышено добавлением химических компонент, таких как ингибиторы ДНК-метилтрансферазы, гистон деацителазы, киназы митоген-активированного протеина (MAPK) и киназы-3 гликоген синтазы (GSK3) [421-423]. Хотя ИПСК имеют такую же морфологию и экспрессируют молекулярные маркеры, похожие на ЭСК, их способность и степень вклада в химеризм сильно варьируют [422, 424]. Таким образом, ИПСК не полностью воспроизводят свойства эмбриональных стволовых клеток [425]; отмечается различие в качестве различных линий ИПСК. Следовательно, другие факторы в дополнение к основным требованиям набора репрограммирования OSK (Oct4, Sox2 и Klf4) могут улучшить качество ИПСК, определяемое их свойствами компетентности зародышевой линии - т.е. свойствами формирования органов и целого организма.

Анализ ChIP секвенирования был выполнен, как описано в предыдущих разделах данной главы. Сайты связывания Tbx3 были определены с помощью компьютерной программы MACS, используя данные специфического ChIP секвенирования Tbx3 и контрольного секвенирования из тех же ЭСК мыши, как описано ранее. Расчет совместной локализации сайтов связывания в геноме мыши был выполнен, как описано в предыдущей главе и в статье [3].

Для более детального исследования того, как Tbx3 может улучшать качество ИПСК, на оборудовании Solexa был выполнен эксперимент ChIP-seq определения связывания и прямых регуляторных мишеней Tbx3 в ЭСК мыши.

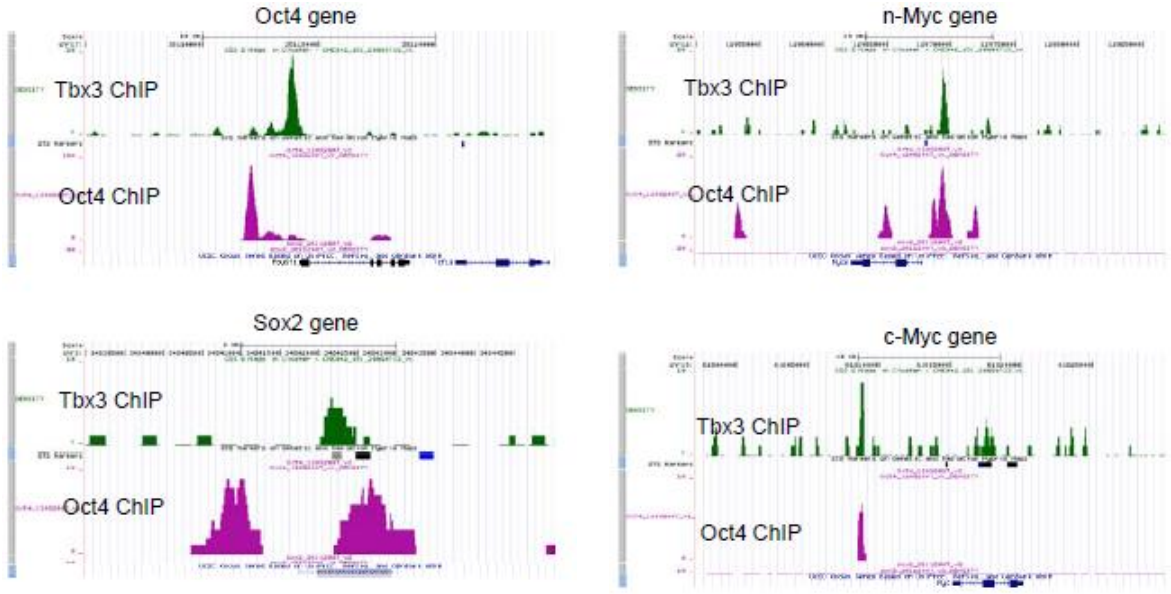


Рис. 3.23. Профили и пики связывания ChIP-seq факторов Oct4 и Tbx3 в районах генов плюрипотентности Pou5f1 (Oct4) и Sox2 (левая панель) и репрограммирования - n-Myc и c-Myc (правая панель) [41].

Кластеризация Tbx3 с ранее картированными в ЭСК мыши ССТФ [3] показала, что Tbx3 имеет общие сайты связывания с классическими ассоциированными с плюрипотентностью ТФ Oct4, Sox2, Nanog и Smad1 (см. рисунок). Tbx3 также имеет гены-мишени (по связыванию сайтов в геномных районах соответствующих генов) для Oct4, Sox2, Sall4,Lefty1, Lefty2 и Zfp42, а также факторов репрограммирования Klf2, Klf4, Klf5, N-myc (также известного как Mycn) и c-myc (Myc)

3.9. Факторы репрограммирования и плюрипотентности

Коды транскрипционных факторов репрограммирования для индукции плюрипотентности. Несмотря на видоспецифичные различия взаимодействий ключевых транскрипционных факторов в геноме, определенные факторы транскрипции

ЭСК могут иметь доминирующий эффект на ассоциированную с плюрипотентностью клеточную идентичность как у мыши, так и у человека. В 2006 году выдающееся исследование Яманака и соавторов продемонстрировало конверсию соматических клеток грызунов в плюрипотентные клетки с помощью ретровирусной трансдукции четырех транскрипционных факторов: Oct4, Sox2, Klf4, c-Myc [406]. Эти репрограммированные клетки, также известные как индуцированные плюрипотентные стволовые клетки (ИПСК), очень близки к ЭСК в терминах морфологии, экспрессии генов, эпигенетических маркеров [19, 406, 431]. Интересно отметить, что тот же набор транскрипционных факторов может индуцировать плюрипотентный фенотип для соматических клеток человека [412]. Способность Oct4 и Sox2 репрограммировать клетки к плюрипотентному состоянию не является неожиданной, принимая во внимание тот факт, что эти два транскрипционных фактора были ранее хорошо изучены как поддерживающие самообновление и плюрипотентность ЭСК [415, 419, 427]. Похожим образом, ТФ c-Myc также был вовлечен в поддержание ЭСК, где было отмечено действие белка Myc как эффектора пути передачи сигнала LIF-Stat3 [445]. В то же время ТФ Klf4 был неожиданным добавлением к коктейлю репрограммирования, поскольку было известно мало информации о белках семейства Klf (Kruppel-like transcription factors) в контексте ЭСК. Тем не менее, следуя открытию Яманака, Нг и соавторы установили, что Klf4, также как и близкие члены того же семейства ТФ Klf2 и Klf5, важен для самообновления ЭСК мыши [436]. Исследование [438] показало, что клеточные пути Oct4 и LIF-Stat3 активируют Klf2 и Klf5, соответственно, для поддержания самообновления ЭСК. Хотя все три Klf белка вовлечены в самообновление ЭСК, фактически есть избыточность в этих трех белках Klf, поскольку только тройной нокаут Klf2, Klf4 и Klf5 в ЭСК мыши индуцирует явный дифференцированный фенотип [436]. Показано, что Klf2 и Klf5 могут заменять Klf4 в репрограммировании соматических клеток [439]. Интересно отметить, что кроме способности генов из тех же семейств, что и Klf4, Sox2, c-Myc заменять их аналогов в репрограммировании [439], несколько факторов Яманака могут быть замещены другими неродственными транскрипционными факторами [40, 424]. Например, Esrrb, орфанный (т.е. не имеющий лигандов) ядерный рецептор, может заменить Klf4 в репрограммировании мышиных эмбриональных фибробласт [424]. Интересно, что другой ядерный рецептор, Nr5a2, может заменить эндогенный Oct4 в репрограмировании соматических клеток грызунов [52].

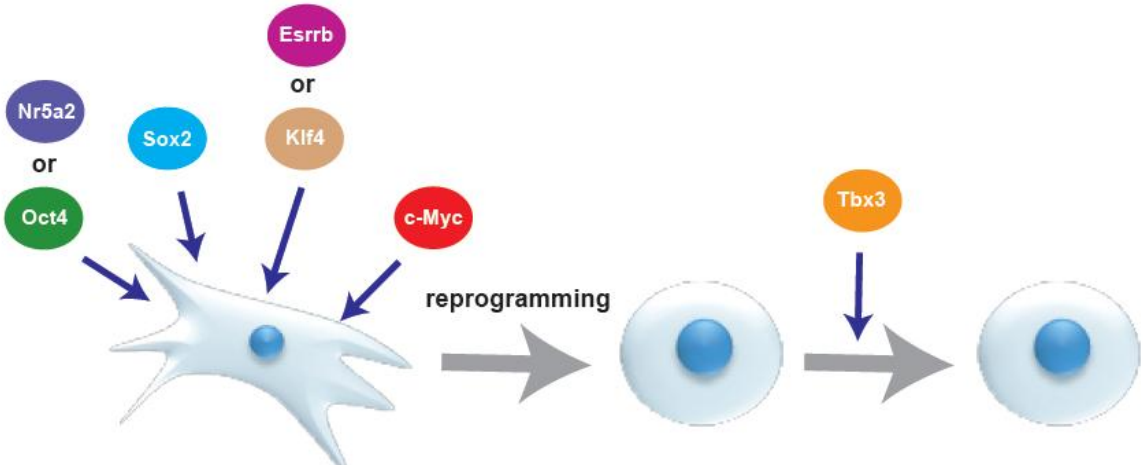


Рис. 3.24. Роль транскрипционных факторов Nr5a2, Esrrb и Tbx3 в репрограммировании. Соматические клетки грызунов могут быть репрограммированы в ИПСК посредством определенного «коктейля» факторов репрограммирования, состоящего из Oct4, Sox2, Klf4 и c-Myc [406]. Интересно, что неродственные транскрипционные факторы могут заменить факторы Яманака в конверсии соматических клеток в плюрипотентные клетки. Например, ядерные рецепторы Nr5a2 и Esrrb могут заменить экзогенные Oct4 и Klf4, соответственно. Более того, Tbx3 (T-box factor), может улучшить качество сгенерированных ИПСК усиливая комптененцию зародышевой линии ИПСК (т.е. способность формировать организм) [52].

Эти результаты добавляют новые перспективы в код репрограммирования плюрипотентного состояния клеток, потому что даже Oct1 и Oct6, близкие члены семейства белков Oct4, неспособны заменить Oct4 в репрограммировании [439]. Репертуар транскрипционных факторов, ассоциированных с репрограммированием, был далее увеличен с открытием свойств Tbx3, который способен значительно увеличить компетенцию зародышевой линии ИПСК грызунов (рис. 3.28) [41]. Примечательно, что полногеномный анализ сайтов связывания Nr5a2, Esrrb и Tbx3, которые как показано в представленной серии работ, вовлечены в репрограммирование, показал что эти сайты имеют тенденцию к совместной локализации в кластерах сайтов Nanog-Oct4-Sox2. Это наблюдение говорит от том, что эти факторы репрограммирования имеют значимую роль в поддержании ЭСК. Такое наблюдение было поддержано экспериментом по потере функции через интерференцию РНК в ЭСК мыши, идентифицировавшим Tbx3 и Esrrb как важные факторы в поддержании самообновления [440]. Кроме того, ранее было показано, что Esrrb важен в поддержании плюрипотентности ЭСК мыши [429].

Ядерный рецептор Nr5a2, связывающийся с проксимальным промотором и проксимальным энхансером гена *Pou5f1* [449], также вовлечен в поддержание ЭСК мыши [449-451]. Принимая во внимание двойную роль транскрипционных факторов в контексте репрограммирования и поддержания плюрипотентности в ЭСК, в процессах дедифференциации и плюрипотентности, показывает, что они должны дополнять друг друга. Тем не менее, заметим, что недавно представленные факторы репрограммирования Esrrb, Nr5a2 и Tbx3 не имеют доказанной роли в поддержании самообновления и плюрипотентности в ЭСК человека. Это может быть связано с видоспецифичными различиями в сети транскрипционных взаимодействий для ЭСК мыши и человека.

Анализ ChIP-seq данных связывания ТФ Eset

Определение пиков ChIP-seq для ТФ Eset было выполнено с помощью программы MACS [294] с пороговым уровнем значимости 1e-12. Было определено 4,633 пиков связывания [39]. Для определения районов маркера модификации гистонов H3K9me3 изменяющих состояние при нокдауне Eset (после РНК-интерференции гена *Eset*), использовалась программа CCAT [514] (переданная автору разработчиком Han Xu). Программа CCAT подходит для анализа районов, обогащенных модификациями гистонов (при работе с параметром “region mode”), определяя более широкие геномные районы, чем локализованные пики, детектируемые большинством программ определения пиков. Получен список Eset-зависимых геномных районов модификации гистонов H3K9me3, ранжированный по разнице (отношению сигнала) высоты профиля секвенирования в клетках между контролем и нокаутом гена *Eset* (корректированный на глубину секвенирования). При пороговом уровне отношения профилей ChIP-seq 2.5 раза, было получено 10,798 Eset-зависимых районов метилирования гистонов H3K9me3.

Для определения генов-мишеней были подсчитаны все гены RefSeq имеющие, по крайней мере, один ChIP-seq пик в окрестности +/-50Кб от старта транскрипции гена. Число генов-мишеней в таком определении составляет 2353 для ССТФ Eset и 4169 для районов H3K9me3 (См. Таблицу).

Для того, чтобы определить основной набор («кор») генов, регулируемых Eset, из построенных списков генов были выбраны все гены RefSeq, которые имели, по меньшей мере, один пик связывания Eset и Eset-зависимый район, обогащенный маркером модификации гистонов H3K9me3. Так сравнение дало список 1283 генов. Если сайты связывания Eset не перекрывались непосредственно с районами H3K9me3,

но находились в промоторном районе одного и того же гена, этот ген рассматривался как ген-мишень. Такое общее определение гена-мишени достаточно обосновано, поскольку прямое перекрывание ChIP-seq сайтов Eset и H3K9me3 дает 1890 «коровых» сайтов и почти столько же генов-мишеней (1171 против 1283 используя определение перекрывания в промоторном районе гена).

Таблица 3.8

Число пиков и генов-мишеней Eset и модификации хроматина H3K9me3

	Eset	H3K9me3	Eset+H3K9me3
Пики ChIP-seq	4633	10798	1890
Гены-мишени	2353	4169	1283

Анализ совместного связывания Eset с другими факторами в геноме мыши. Был выполнен анализ перекрывания 4,633 пиков Eset с другими транскрипционными факторами (Oct4, Sox2, Nanog, Suz12 и др.) в ЭСК мыши с помощью расчета общих пиков в геноме по сравнению с данными предыдущей работы [3]. Допускалось до 200 нт между границами двух пиков, что соответствует точности разрешения метода (длине фрагмента ChIP-seq в экспериментах). Для сравнения полученного числа перекрываний (общих районов пиков) в геноме с ожидаемым по случайным причинам, использовались данные контрольной библиотеки секвенирования ДНК без иммунопреципитации. Используя заниженный порог определения пиков (1e-3) в программе MACS для контрольной библиотеки было определено 40,000 случайных районов генома. Такой подход позволяет скорректировать возможную неравномерность при фрагментации и картировании прочтений ChIP-seq. Было рассчитано число пересечений этих случайных пиков с сайтами связывания каждого исследованного ТФ, а также Eset. Также было рассчитано число пересечений набора сайтов Eset с сайтами каждого исследованного ТФ. Статистическая значимость обогащения числа пересечений геномных координат ТФ и Eset была рассчитана с помощью точного критерия Фишера, была показана ассоциация с сайтами факторов плюрипотентности [39].

3.10. Сайты связывания в геноме в зависимости от дозового эффекта и взаимодействия ко-факторов на примере ССТФ Smad2 в ЭСК мыши

Дозовый эффект связывания ТФ. Исследование эффекта связывания ТФ и изменения экспрессии генов на микрочипах выполнялось для ССТФ Smad2 в ЭСК мыши.

Нодал и Активин (Nodal, Activin) являются морфогенами суперсемейства сигнальных молекул TGFbeta, определяющих направление дифференцировки клеток в доз-зависимой и зависящей от расстояния манере. Во время раннего эмбрионального развития путь передачи сигнала Nodal/Activin отвечает за спецификацию мезодермы, эндодермы и мезендодермы. В противоположность этого действия по направлению клеточной дифференциации, этот путь играет важную роль в поддержании самообновления и плюрипотентности в ЭСК и в клетках эпибласта. Молекулярные основы восприятия клеткой градиентов сигнала Nodal/Activin и принятия отдельной клеткой направления дальнейшего развития или дифференцировки остаются недостаточно изученными. Был выполнен эксперимент по нарушению уровней эндогенного сигнала в ЭСК мыши ведущий к выходу клеток из состояния самообновления в расходящиеся программы дифференциации клеток.

Увеличение сигналов Nodal выше базального уровня прямой стимуляцией активина направляет дифференциацию по направлению к мезендодермальным линиям (lineages) тогда как подавление (репрессирование) передачи сигнала специфичным ингибитором рецептора Nodal/Activin, имеющим название SB431542, индуцирует дифференциацию в трофектодерму.

Изменение уровня сигнала Nodal/Activin ведет к транскрипционной регуляции специфических групп генов-мишеней. Цветом выделены имена генов маркеров для эндодермы(endoderm), мезодермы(mesoderm), трофектодермы (green) и онкогены. Измерение в шкале log2 по средним контролям в n = 4 биологических репликах.

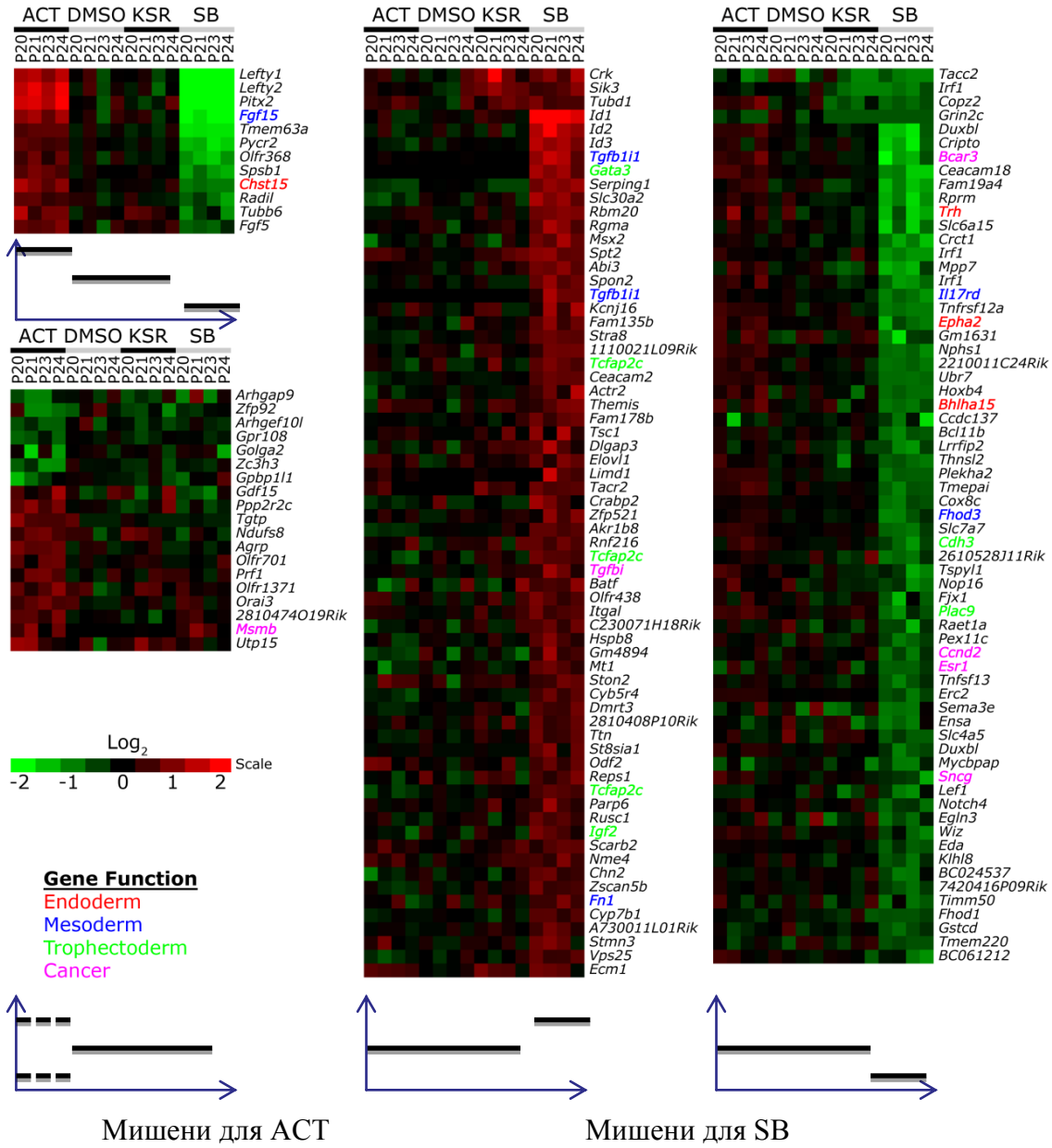


Рис. 3.25. Тепловая карта экспрессии на микрочипах в ЭСК мыши после обработки активатором Activin (ACT), DMSO или KSR (контроли) и ингибитором SB. Красный цвет – повышенная экспрессия, зеленый – пониженная экспрессия (Lee et al., 2011).

Увеличение/уменьшение уровня сигнала нодал/активин (Nodal/Activin) ведет к транскрипционной регуляции специфических групп мишеней Smad2 в ЭСК (эмбриональных стволовых клетках) мыши.

Для анализа того, как количественный сигнал Nodal/Activin качественно транслируется в принятие решения о судьбе клетки, был выполнен эксперимент иммунопреципитации хроматина для фосфорилированной формы ТФ Smad2,

первичного транскрипционного фактора в пути передачи сигнала Nodal/Activin. Профили связывания ChIP-seq представлены на рисунке.

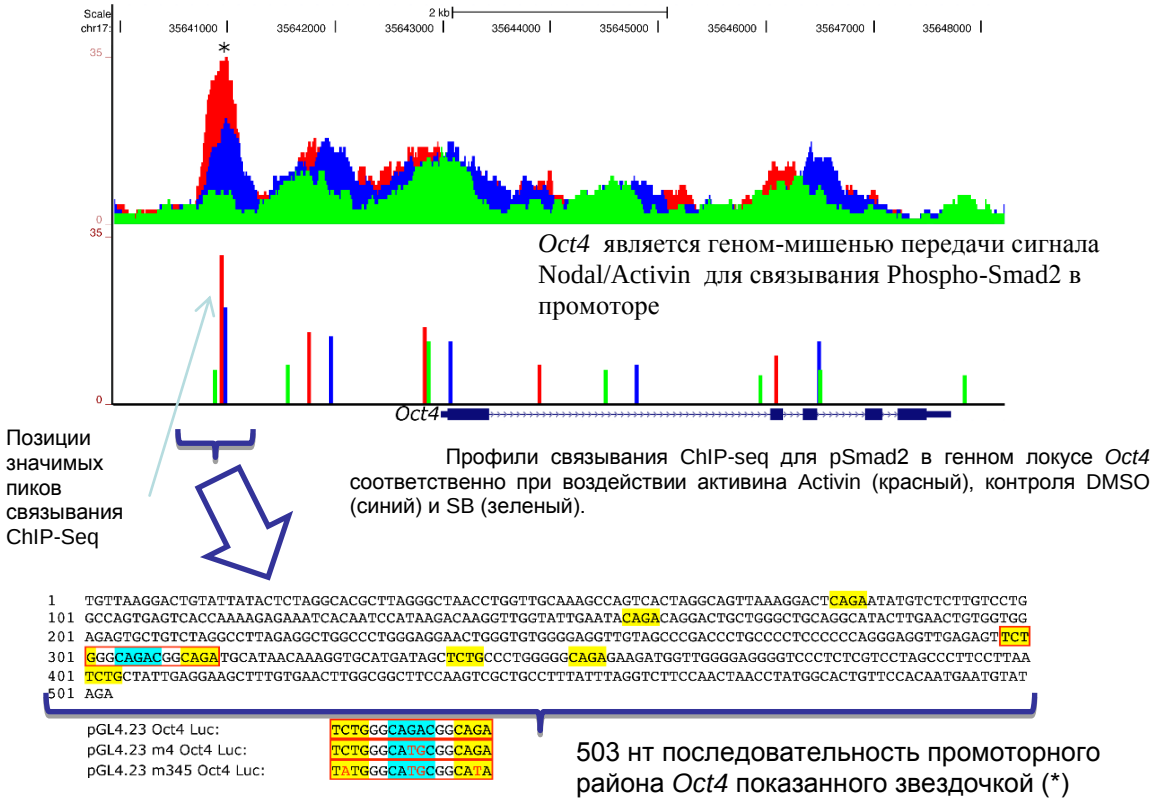


Рис. 3.26. Профили связывания Smad2 в трех условиях и поиск мотивов ССТФ в пике ChIP-seq.

Было показано, что Smad2 связывается и регулирует различные подмножества генов-мишеней в доз-зависимой манере. Исключительно важно, что сигнал Nodal/Activin непосредственно контролирует основной регулятор плюрипотентности Oct4 через градации связывания Smad2 в его промоторном районе. Следовательно, стволовые клетки интерпретируют и выполняют сигнальные инструкции Nodal/Activin через соответствующий градиент фосфорилирования Smad2, который селективно задает уровень самообновления против альтернативной программы дифференциации клеток посредством прямого регулирования различных генов-мишеней и экспрессии Oct4.

3.11. Геномные карты сайтов связывания ТФ для генома человека

Построены геномные карты связывания ТФ c-Myc, ER, FOXA1 и выполнена компьютерная интеграция данных о сайтах связывания этих ТФ с микрочиповыми данными экспрессии генов в геноме человека.

Для эмбриональных стволовых клеток человека показан тот же эффект ко-локализации ключевых факторов плюрипотентности, включающих OCT4-NANOG-SOX2, что и для ЭСК мыши (рис. 3.31). При этом использовались данные экспериментов ChIP-seq из серии работ, выполненных в Геномном институте Сингапура.

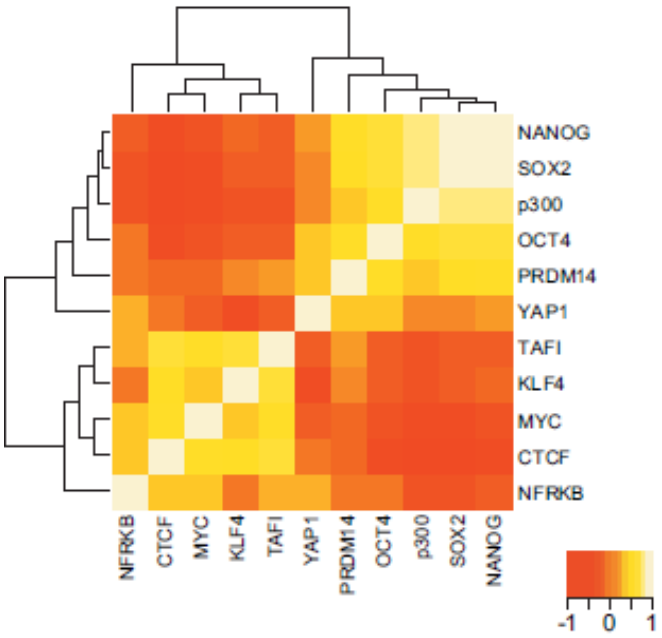


Рис. 3.27. Ко-локализация ССТФ в геноме человека, включая NFRKB и PRDM14 [42].

Исследование нового фактора плюрипотентности – PRDM14 – было выполнено с помощью параллельного скрининга нокаутов генов человека в культуре эмбриональных стволовых клеток человека H1. Определение ССТФ PRDM14 в геноме человека позволили определить de novo мотив связывания (рис. 3.32). Анализ совместного распределения сайтов связывания нескольких ТФ в геноме человека показал локализацию сайтов PRDM14 с регуляторными районами плюрипотентности, формируемыми сайтами связывания ТФ OCT4, SOX2, NANOG.

Исследование нового фактора плюрипотентности – PRDM14 – было выполнено с помощью параллельного скрининга нокаутов генов человека в культуре эмбриональных стволовых клеток человека H1 [42]. Схема эксперимента состояла в следующем. Была подготовлена репортерная конструкция, содержащая флюоресцирующий белок GFP, находящийся под контролем промотора гена POU5F1. Такая конструкция была введена в клеточную линию H1 ЭСК человека. Недифференцированное состояние стволовых клеток детектировалось по экспрессии маркера плюрипотентности OCT4 (гена *POU5F1*) соответствующего высокой интенсивности флюоресценции клеток. Нарушения такого состояния, вызванное нокаутом отдельных генов посредством РНК интерференции (RNAi) из-за добавления малых интерферирующих РНК (siRNA), приводило к потере флюоресценции.

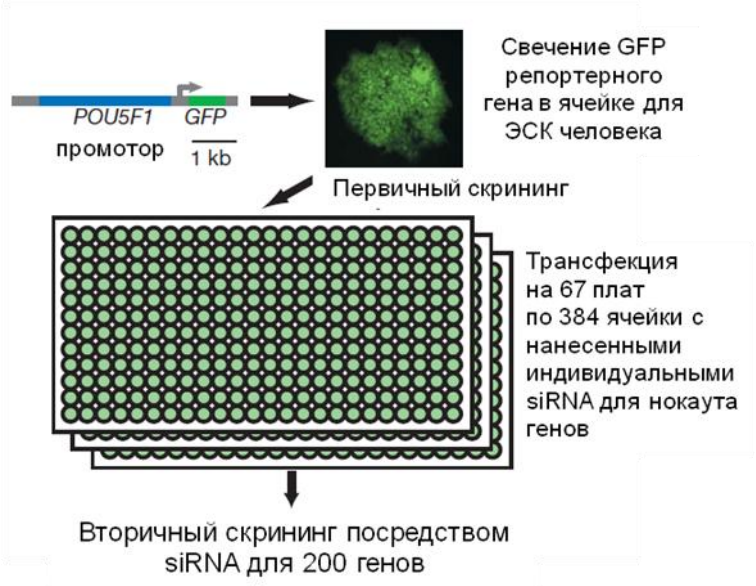


Рис. 3.28. Определение мотива de novo для ССТФ PRDM14 в геноме человека [42].

Был выполнен массовый эксперимент скрининга флюоресценции клеток с внедренной конструкцией GFP под промотором Oct4 и индивидуальной интерферирующей РНК (siRNA), подавляющей экспрессию генов для каждого из 21,121 генов человека. Численные значения флюоресценции ячеек, соответствующих нокаутированным генам, после нормализации на число выживших клеток в ячейке, позволили составить таблицу сохранения свойств плюрипотентности для исследованного набора генов.

Кроме подтверждения важности самого гена *POU5F1*, существенного для поддержания свойств ЭСК, среди верхних 5% генов из списка были найдены HCFC1, TCL1A, ZSCAN10, ZIC3, NANOG и ZNF143, существенные для ЭСК мыши [42]. Среди

новых генов был найден PRDM14 – транскрипционный фактор, регулирующий экспрессию OCT4. Далее для некоторых факторов, включая PRDM14, был выполнен эксперимент ChIP-seq по определению сайтов связывания в геноме.

Для эмбриональных стволовых клеток человека показан тот же эффект ко-локализации ключевых факторов плюрипотентности, включающих OCT4-NANOG-SOX2, что и для ЭСК мыши [42, 52].

Анализ совместного распределения сайтов связывания нескольких ТФ в геноме человека показал локализацию сайтов PRDM14 с регуляторными районами плюрипотентности, формируемыми сайтами связывания ТФ OCT4, SOX2, NANOG в масштабе генома (рис. 3.29). Тепловая карта ко-локализации строилась тем же методом, что и для сайтов связывания в геноме мыши.

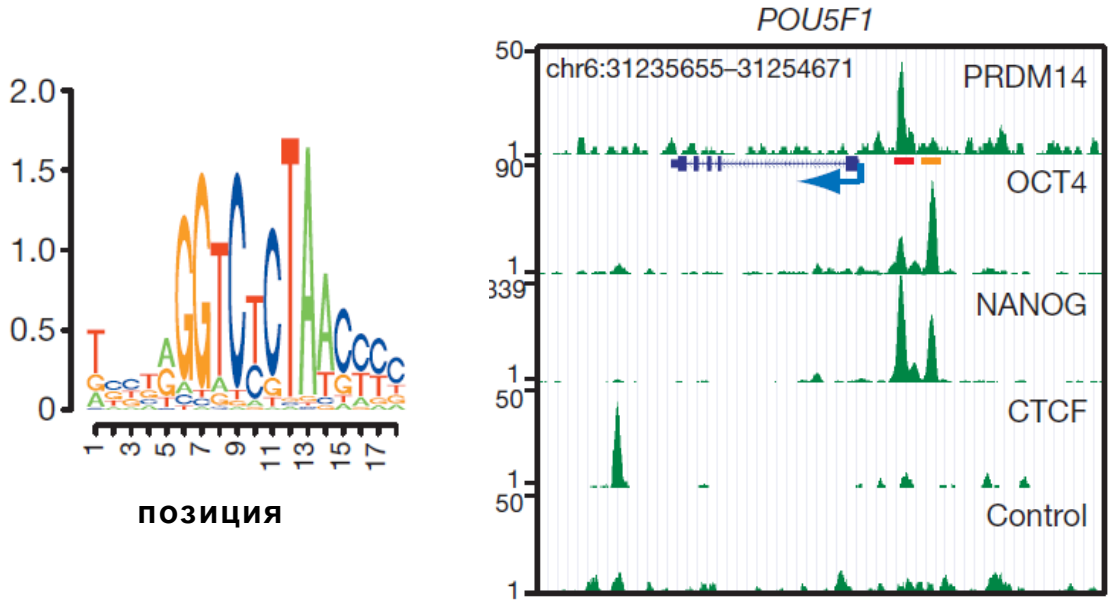


Рис. 3.29. (Левая панель) Определение мотива de novo для ССТФ PRDM14 в геноме человека.
(Правая панель) Расположение пиков связывания транскрипционных факторов PRDM14, OCT4, NANOG, CTCF в локусе гена *POU5F1* для профилей ChIP-seq [42].

Анализ нуклеотидных последовательностей сайтов связывания PRDM14 в геноме человека впервые позволил определить мотив связывания (рис. 3.29, левая панель). Правая панель рисунка 3.29 показывает совместное расположение сайтов связывания PRDM14, OCT4, NANOG в промоторе гена *POU5F1*, который является мишенью их воздействия. При этом CTCF не имеет специфического связывания в промоторе, и контрольное секвенирование ДНК без иммунопреципитации не дает никаких пиков.

Рисунок 3.30 показывает сравнительные эффекты активации ТФ плюрипотентности в ЭСК мыши и человека [40]: выделены консервативные регуляторные районы и вариабельные (специфические для генома человека) регуляторные мишени для OCT4 и NANOG.

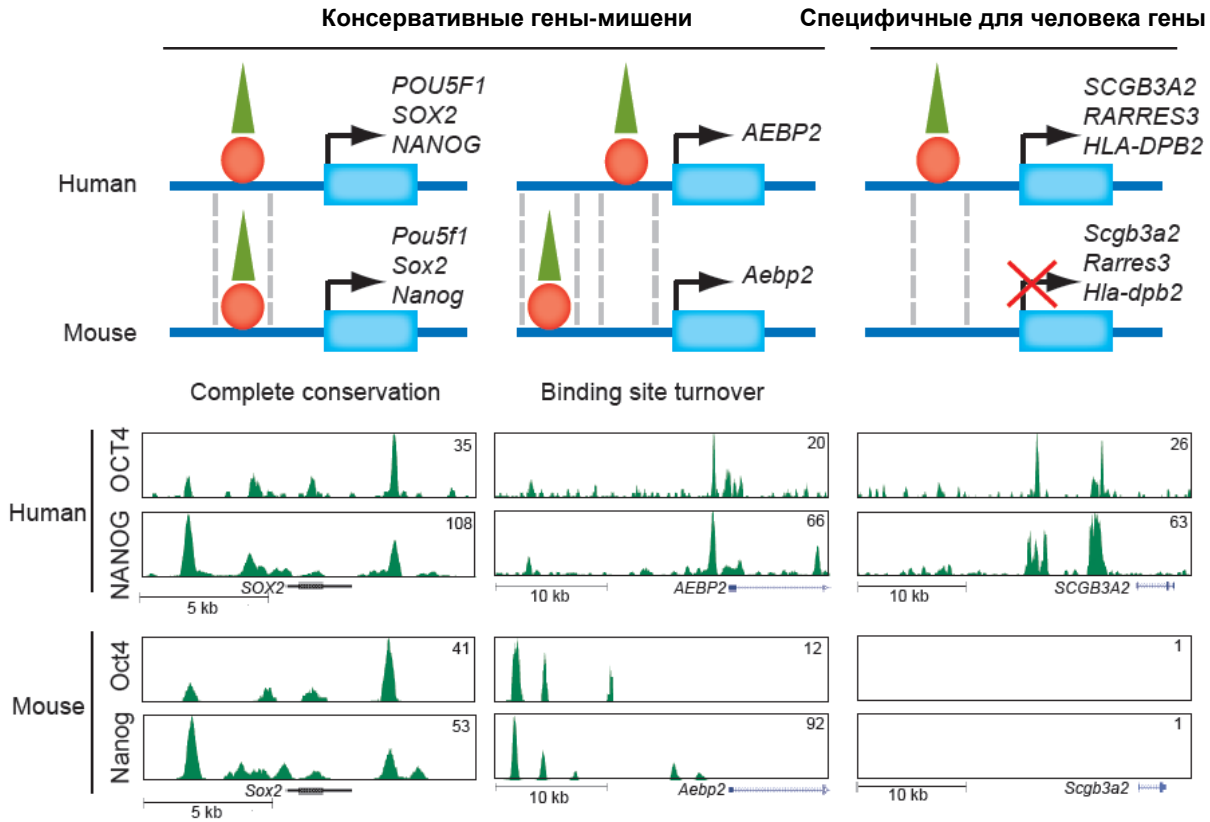


Рис. 3.30. Эффекты активации ТФ плюрипотентности в ЭСК мыши и человека [40].

Таким образом, в данной главе описаны полногеномные карты сайтов связывания транскрипционных факторов в эмбриональных стволовых клеток, полученные по экспериментальным данным ChIP-seq в геноме человека, а также в геноме мыши, показаны тепловые карты сближенности сайтов, описаны их кластеры и нуклеотидные мотивы [3, 39-42, 52, 54].

Заключение к Главе 3

Представлено описание и анализ полногеномных карт сайтов связывания транскрипционных факторов, полученных по экспериментальным данным ChIP-seq в геномах человека и мыши [3, 39-42, 52]. Исследовалось распределение ССТФ онкогенов в геноме человека [9, 13]. Основные разделы посвящены исследованию факторов плюрипотентности в стволовых клетках. Подробно показаны эксперименты на мыши, показано применение разработанного подхода для ЭСК человека.

В данной главе обоснованы следующие положения выносимые, на защиту:

- Полногеномные карты сайтов связывания транскрипционных факторов в эмбриональных стволовых клетках, построенные по данным ChIP-seq для c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши свидетельствуют о совместной локализации групп сайтов связывания транскрипционных факторов Oct4, Sox2, Nanog, с одной стороны, и c-Myc, n-Myc с другой.

- Нуклеотидные последовательности, окружающие сайты связывания транскрипционного фактора Smad2 в геноме мыши, содержат специфические группы нуклеотидных мотивов, соответствующих потенциальным сайтам связывания других транскрипционных факторов. Эти мотивы различаются для сайтов связывания Smad2, найденных в эмбриональных стволовых клетках мыши при действии внешних факторов - белка Activin и ингибитора SB431542, соответственно.

Для экспериментов по секвенированию ДНК сопряженному с иммунопреципитацией ChIP-PET и ChIP-seq разработаны компьютерные программы обработки данных, выделения статистически значимых пиков. С их помощью были проанализированы исходные данные секвенирования и определены сайты связывания белков c-Myc, STAT1, FOXA1, ERα, PRDM14 в геноме человека [9, 13], транскрипционных факторов и регуляторов Nanog, Oct4, Sox2, Klf4, E2f1, Esrrb, CTCF, n-Myc, c-Myc, Smad1, STAT3, Tcfcf21l, Zfx, Suz12 [3].

Проведен компьютерный анализ данных экспериментов ChIP-PET и ChIP-seq для транскрипционных факторов MYC и ERα, соответственно, в культурах клеток опухолей человека P493 и MCF-7. На основе компьютерной обработки этих экспериментальных данных установлена положительная корреляция ($p<0.001$) между силой связывания транскрипционных факторов ERα и MYC, измеренной с помощью кПЦР, и числом прочтений ДНК в экспериментах иммунопреципитации хроматина.

Выявлены нуклеотидные мотивы транскрипционных факторов, связывающихся в окрестностях сайтов ERα.

Выполненная работа позволяет аргументировать следующие выводы:

- С помощью компьютерного анализа данных экспериментов ChIP-seq на эмбриональных стволовых клетках мыши впервые построена термокарта совместной локализации транскрипционных факторов Oct4, Nanog, Sox2, Klf4, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши. Показана совместная геномная локализация сайтов связывания транскрипционных факторов Oct4, Nanog, Sox2 и Klf4, относящихся к ключевым регуляторам плюрипотентности.

- Впервые по данным экспериментов ChIP-seq на эмбриональных стволовых клетках мыши определены группы сайтов связывания транскрипционного фактора Smad2 в условиях активации и подавления экспрессии гена Smad2 под действием внешних факторов - белка Activin и ингибитора SB431542, соответственно. В геномном окружении сайтов Smad2 найдены специфичные группы нуклеотидных мотивов, соответствующих потенциальным сайтам связывания других транскрипционных факторов.

- На основе компьютерной модели эксперимента с последовательным подавлением транскрипции генов в эмбриональных стволовых клетках (ЭСК) человека показана роль транскрипционного фактора PRDM14 в поддержании плюрипотентности. Для транскрипционного фактора PRDM14 по данным ChIP-seq найдены его гены-мишени в ЭСК человека, включающие OCT4. Впервые определена структура сайта связывания PRDM14.

- С помощью компьютерного анализа данных экспериментов ChIP-seq в ЭСК человека построена термокарта расположения кластеров сайтов связывания для транскрипционных факторов OCT4, NANOG, SOX2 и PRDM14 в геноме человека. Показано совместное геномное расположение сайтов связывания транскрипционных факторов OCT4, NANOG, SOX2 в ЭСК человека, аналогичное расположению сайтов связывания их гомологов в ЭСК мыши.

Глава 4. МОДИФИКАЦИИ ХРОМАТИНА И СВЯЗЫВАНИЕ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ В ГЕНОМЕ

4.1. Введение к Главе 4.

Данная Глава посвящена анализу упаковки хроматина в геноме, выявленному с помощью методов секвенирования, и регуляции экспрессии генов в зависимости от состояния хроматина и условий доступности сайтов связывания транскрипционных факторов.

В данной Главе решается следующая научная задача:

Компьютерное исследование ассоциации сайтов связывания транскрипционного фактора ERα с определенными с помощью технологии ChIP-seq маркерами хроматина, в частности, модификациями гистона H3 (H3K4me3, H3K4me1, H3K27me3, H3K9me3, H3K9ac, H3K14ac), и создание метода предсказания сайтов связывания транскрипционного фактора ERα в геноме человека на основе профилей модификаций гистонов.

Глава содержит разделы, посвященные исследованию ассоциаций сайтов связывания ТФ с модификациями хроматина, предсказанию сайтов связывания в геноме человека с помощью компьютерной модели, учитывающей состояние хроматина, определенное по данным ChIP-seq. Рассмотрена общая проблема предсказания сайтов связывания и регуляции транскрипции на основе модификаций хроматина и открытости ДНК, оцененной по присутствию нуклеосом. Приведены примеры применения анализа связи присутствия нуклеосом в промоторных районах генов, используя профиль присутствия нуклеосом, определенный в экспериментах по полногеномному секвенированию для дрожжей [50, 51, 62].

Представлен анализ нуклеотидных последовательностей сайтов формирования нуклеосом, полученных в результате полногеномного секвенирования в связи с анализом регуляторных районов транскрипции генов эукариот [51, 62, 63].

Результаты работ по анализу ассоциации сайтов связывания ТФ ERα с модификациями хроматина, в частности гистона H3 (H3K4me3, H3K27me3, H3K4me1), определенными с помощью технологии ChIP-seq, и работ по созданию метода предсказания сайтов ERα в геноме на основе профилей модификаций гистонов опубликованы в работах автора [13, 21, 37].

Показана возможность предсказания сайтов связывания транскрипционных факторов в геноме по данным ChIP-seq модификаций гистонов [13].

Дано обсуждение общей проблемы формирования структур нуклеосом, контролирующих доступность ДНК для связывания транскрипционных факторов в промоторных участках генов, и регуляции транскрипции.

4.2. Исследование нуклеосомной упаковки и расположения сайтов связывания транскрипционных факторов в геноме дрожжей

Геномная ДНК в ядре клетке большей частью находится не в свободном состоянии, а связана с белками, в основном в форме нуклеосом. Остающаяся часть фракции белков состоит в основном из ассоциированных с хроматином факторов, включая ферменты модификации гистонов и ДНК, белков, способствующих передвижению нуклеосом, и специфичных ДНК-связывающих белков - транскрипционных факторов. Понимание регуляции экспрессии генов требует детального исследования процессов взаимодействия перечисленных групп белков - их кооперации или соревнования за доступ к двухцепочечной геномной ДНК. Полногеномные исследования с помощью экспериментов иммунопреципитации на микрочипах (ChIP-chip) позволили получить картину полногеномного распределения гистонов в геноме дрожжей [275-277].

Ранее была показана недопредставленность нуклеосом в промоторных районах генов. Геном дрожжей в целом очень хорошо аннотирован, содержит порядка шести тысяч генов, большей частью состоит из белок-кодирующих районов, межгенные участки и интроны занимают незначительную часть генома, поэтому полногеномное исследование распределения нуклеосом выполнить технически гораздо проще, чем для геномов млекопитающих. Кроме того, из-за большой плотности расположения генов дистальные регуляторные районы переходят в область соседних генов. Было показано, что транскрипционные факторы связываются в основном в промоторных районах, а не в транскрибируемых районах. Представленность нуклеосом отличается между промоторными районами и такая разница в присутствии нуклеосом коррелирует с вероятностью связывания с ДНК транскрипционных факторов [302].

Несомненен общий эффект связи открытого состояния упаковки ДНК (нуклеосомной упаковки, или состояния хроматина) в промоторных районах генов с присутствием белковых факторов транскрипции и активацией работы генов. Однако, количественное соотношение состояния хроматина и повышение экспрессии гена, соотношение между связыванием различных транскрипционных факторов, имеющих различную пространственную структуру укладки и различные функции в клетке, требует более детального исследования.

Соревнование за связывание между белками нуклеосом (гистонами) и белками – транскрипционными факторами является следствием того, что все эти белки имеют предопределенную структурой способность связываться к одному и тому же сайту ДНК. Некоторые транскрипционные факторы способны связываться с ДНК на внешней поверхности нуклеосомы, но стерическая закрытость части двойной цепи ДНК и сильный изгиб ДНК вокруг нуклеосомы не дает большинству факторов связываться с ДНК в составе нуклеосом. Нуклеосомы могут опосредовать взаимодействия между транскрипционными факторами пассивным образом, как посредники в соревновании белков за связывания с ДНК. Например, при расположении рядом двух сайтов связывания двух разных факторов, сайт связывания одного из факторов будет иметь меньшую представленность нуклеосомы, чем при отсутствии второго фактора, поскольку нуклеосомная конфигурация не позволяет занимать два сайта одновременно. Эта пониженная частота присутствия нуклеосомы приводит к повышению эффективного связывания белка с ДНК в первом сайте. В таком сценарии кооперативное связывание факторов опосредовано не прямыми взаимодействиями между белками, а пассивным присутствием нуклеосомы. Такие эффекты были продемонстрированы экспериментально [303].

Пассивная роль нуклеосом во взаимодействиях транскрипционных факторов – лишь один из путей влияния нуклеосом на связывание транскрипционных факторов в масштабе генома. На связывание влияют свойства нуклеотидной последовательности – предпочтение, или специфичность участка ДНК для контакта с гистоновым октамером и упаковки в структуру нуклеосомы. Нуклеосомы не имеют специфичных контактов между аминокислотной последовательностью и нуклеотидными основаниями ДНК, которые характерны для транскрипционных факторов, но формируют предпочтения к связыванию с ДНК, связанными со способностью молекулы ДНК быть повернутой вокруг нуклеосомы, изгибной жесткостью и соответствующими контекстными характеристиками. Одно из проявлений такого предпочтения к связыванию – 10-нуклеотидная периодичность в распределении определенных динуклеотидов, формирующих структурный изгиб цепи ДНК, таких как AA, TT [264, 276]. Такая периодичность отражает спиральную структуру молекулы ДНК. В недавних работах внимание было привлечено к более длинным A-богатым районам (политрактам), которые являются ингибиторами (запрещающими сигналами) для связывания с нуклеосомой [264, 276, 305]. Практическое исключение этих последовательностей из центральных районов нуклеосом было впервые отмечено в то же время, когда были описаны предпочтения в периодичности динуклеотидов [515], но лишь относительно

недавно стало ясно, что эти политракты вносят существенный вклад в способность предсказания формирования нуклеосом [264, 276, 305].

Каплан и соавторы [264] продемонстрировали, что свойства нуклеотидной последовательности лишь частично определяют специфичность связывания и позиционирование нуклеосомы in vivo. В этой работе, позиции нуклеосом были определены с помощью глубокого секвенирования хроматина клеток дрожжей, разрезанного нуклеазами, а также из in vitro восстановленного хроматина используя гистоновые белки цыпленка и геномную ДНК дрожжей [264]. Было установлено, что контекстные характеристики нуклеотидных последовательностей в составе нуклеосом похожи в нативном хроматине и восстановленном хроматине. Более того, геномные локусы, которые связаны с транскрипционными факторами, имели, в среднем, более низкую представленность нуклеосом (меньшую плотность нуклеосом в геномном районе), чем несвязанные сайты даже когда представленность нуклеосом определялась in vitro. Это позволяет предположить, что внутренние (контекстные) свойства нуклеотидной последовательности имеют некоторый эффект на отбор транскрипционными факторами своих сайтов связывания in vivo. Каплан и соавторы далее утверждали, что нет значительных различий в картировании нуклеосомных позиций, когда хроматин был предварительно зафиксирован формальдегидом по сравнению с более общепринятой процедурой картирования без фиксации [264].

Картирование секвенированных фрагментов без фиксации (перекрестного соединения - crosslinking) предполагает, что сайты, занятые нуклеосомами in vivo остаются связанными с этими сайтами во время эксперимента. Показанная в работе близость данных подтверждает, что в целом это так. Данные [264] были повторно проанализированы в контексте имеющихся данных ChIP-chip [278] вместе с новыми, более точными экспериментами ChIP-qPCR для нескольких сайтов [51]. Подтверждена роль внутренних, свойственных нуклеосоме, предпочтений к связыванию в расположении транскрипционных факторов, хотя в целом анализ скорее показал низкий уровень присутствия нуклеосом в ССТФ, чем количество информации, которое дает такое присутствие. Показано, что роль присутствия нуклеосом состоит не в точном позиционировании, а в определении средней нуклеосомной плотности расположения нуклеосом в районе, соответствующем размеру промотора (600 нт).

Нуклеотидные последовательности генома дрожжей (релиз - август 2008 г.) и файлы характеристик генов были получены из базы данных SGD (Saccharomyces Genome Database). Промоторные районы были определены как 600 нт перед стартом транскрипции белок кодирующих генов. Старты транскрипции были взяты из

аннотации генов SGD. Геномные позиции проб ChIP микрочипа и значений уровней связывания в ChIP-chip эксперименте (значимость наблюдения связывания - *p*-value) с использованием различных помеченных эпитопами транскрипционных факторов в нормальных условиях роста (среда YPD, 30uC) были получены из работы [278]. Карты расположения фрагментов нуклеосомной ДНК, представленные в работе [264] были получены из архива данных секвенирования GEO (номер GSE13622). Координаты картированных последовательностей в геноме дрожжей были конвертированы в координаты релиза SGD версии августа 2008 года. Геномные положения сайтов связывания транскрипционных факторов из ChIP-chip экспериментов были взяты из работы MacIsaac и соавторов. [279]. Для каждого потенциального сайта связывания использовалось значение уровня значимости p-value сигнала связывания ChIP, поставленное в соответствие этому сайту. Мотивы связывания со значением p, не превышающим 0.001, были классифицированы как связанные соответствующим транскрипционным фактором. Всего 41 транскрипционный фактор имел в геноме 50 и более сайтов, связанных по этому критерию, и был использован для дальнейшего анализа. Транскрипционные факторы с меньшим числом сайтов в геноме были исключены из-за недостаточной статистики. Для анализа связывания ТФ Abf1, мотивы Abf1 были разделены на группы, соответствующие значениям p-value значимости связывания в ChIP-chip в интервалах 0.001–0.01, 0.01–0.1, 0.1–0.5, и выше 0.5.

Картирование прочтений нуклеосомных фрагментов на геном.

Данные прочтений нуклеосомной ДНК, представленные в [264] состоят из позиций 5'-концов геномных фрагментов. Зная, что размер нуклеосомы составляет 146 нт, что соответствуют размеру селектированных фрагментов в эксперименте и размеру разрезанию ДНК нуклеазами, прочтение (секвенирование) фрагмента с одной стороны рассматривалось всегда как фрагмент 146 нт. Был построен профиль покрытия такими фрагментами всего генома. Для каждой позиции было рассчитано число - покрытие фрагментами нуклеосомной ДНК, которое являлось мерой присутствия нуклеосомной структуры в данной конкретно взятой точке генома. Для большинства расчетов профилей расположения в геноме, число таких прочтений усреднялось в окне, центр которого располагался в центре исследуемой позиции. Использовались окна размером 15, 40, 75, 150, 300 и 600 нт (см. рис. 4.1 для примера расположения окон для сайтов Abf1).

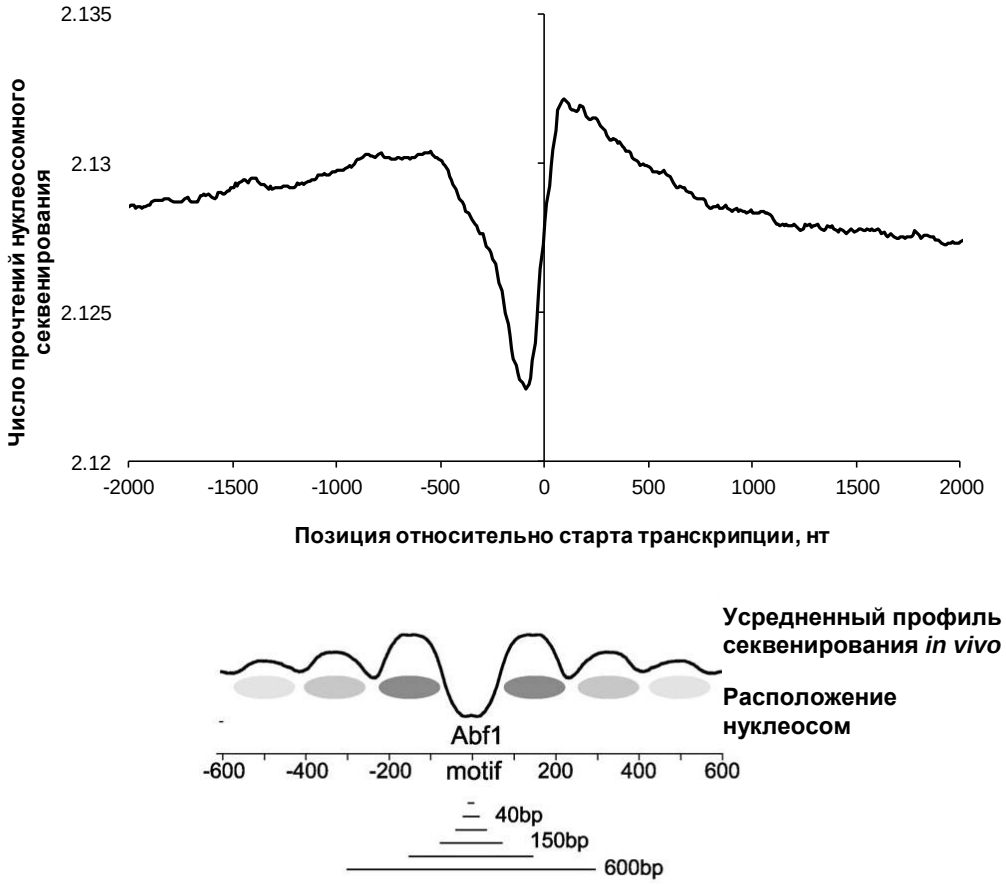


Рис. 4.1. Эффект расположения нуклеосом в промоторных районах вокруг стартов транскрипции (верхняя панель) и вокруг сайтов связывания Abf1 (нижняя панель) по данным секвенирования нуклеосомной ДНК в геноме дрожжей.

Низкая представленность нуклеосом на сайтах связывания транскрипционных факторов может различаться даже на сайтах с низкой представленностью транскрипционных факторов

Для проверки корректности обработки данных полногеномного секвенирования, выполнено сравнение с опубликованными результатами в работе [264].

Был выполнен расчет данных секвенирования нуклеосомной ДНК выполненный в Геномном Институте Сингапура, и представленный автором в GEO (GSE26392). Был рассчитан профиль плотности расположения нуклеосом относительно стартов транскрипции генов дрожжей. Для каждого гена рассчитывался профиль секвенированных фрагментов, данные усреднялись. Подтвержден известный ранее эффект понижения плотности нуклеосом в промоторе и повышения частоты присутствия нуклеосом непосредственно после старта транскрипции (см. Рис. 4.1). Был рассчитан также профиль средний расположения нуклеосом для групп сайтов связывания, в частности для ССТФ Abf1 (см. рис. 4.2). Для сайтов Abf1 расчет

проводился отдельно для групп сайтов связывания с разной интенсивностью, разделенным по уровням значимости от самых сильных до более слабых. Наиболее выраженный профиль имеет глубокий минимум в центре сайта для мотивов связывания с высоким уровнем значимости (значение p для ChIP-chip Abf1 меньше 0.001). Видны локальные максимуму соседних нуклеосом строго позиционированных справа и слева от сайтов. При варьировании строгости определения связывания от (p<0.001) до (p около 0.5 и выше, что формально не являются значимым), наблюдалось отсутствие нуклеосом в центре сайта и присутствие на фланкирующих позициях, причем это различие не исчезало даже для слабых сайтов.

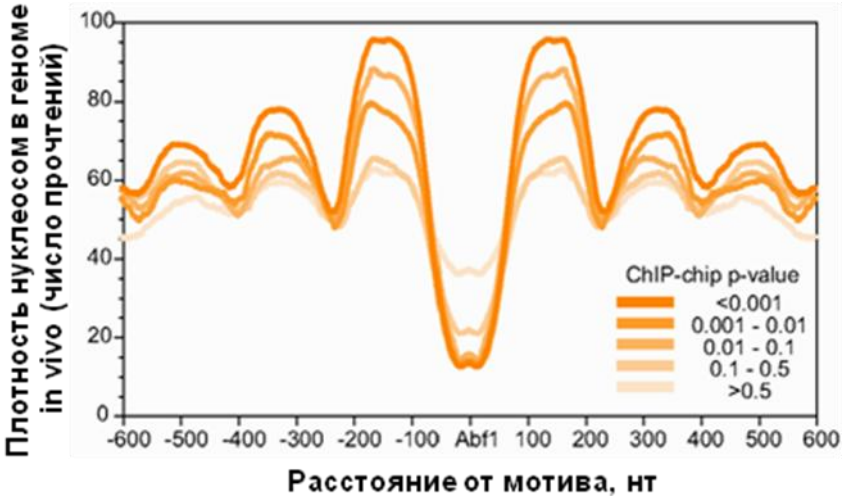


Рис. 4.2. Эффект расположения нуклеосом в промоторных районах (верхняя панель) и вокруг сайтов связывания Abf1 (нижняя панель) по данным секвенирования нуклеосомной ДНК в геноме дрожжей [51].

Для количественной оценки взаимосвязи между низкой плотностью нуклеосом и связыванием ТФ был исследован вопрос, насколько хорошо плотность нуклеосом как численная мера может отличить, связанные сайты от несвязанных случайных сайтов в промоторных районах. Использовалась площадь под кривой ошибок предсказания ROC (ROC AUC: receiver operator characteristic area under the curve) как мера такой количественной ассоциации [516]. Действительно, даже для сайтов Abf1 с низким уровнем значимости, значения площади под кривой ROC AUC превышают уровень, который ожидается по случайным причинам, то есть эффект зависимости есть и предсказание возможно.

Был проведен более детальный анализ для 41 транскрипционного фактора, для каждого из которых было найдено в геноме, по меньшей мере, 50 сайтов, содержащих мотив и связанных по данным ChIP-chip эксперимента, представленного в [278] (с уровнем значимости p<1e-3), и выявленным мотивом в [279].

Для предсказания сайтов (бинарной классификации – связанные сайты – случайные участки промоторных районов) использовался только уровень плотности нуклеосом по данным секвенирования, предполагая, что участки с низкой плотностью нуклеосом содержат сайты, а с высокой плотностью – не содержат. Заметим, что точного разделения получить не удается, но для каждого порогового значения есть свой показатель точности. Для общей оценки точности классификации использовалась площадь под кривой ошибок ROC предсказания сайтов по численному уровню профиля секвенирования нуклеосомной ДНК при варьировании порога. Такие характеристики площади под кривой ошибок были рассчитаны для выборок сайтов каждого из 41 транскрипционных факторов. Результаты представлены на рисунке 4.3.

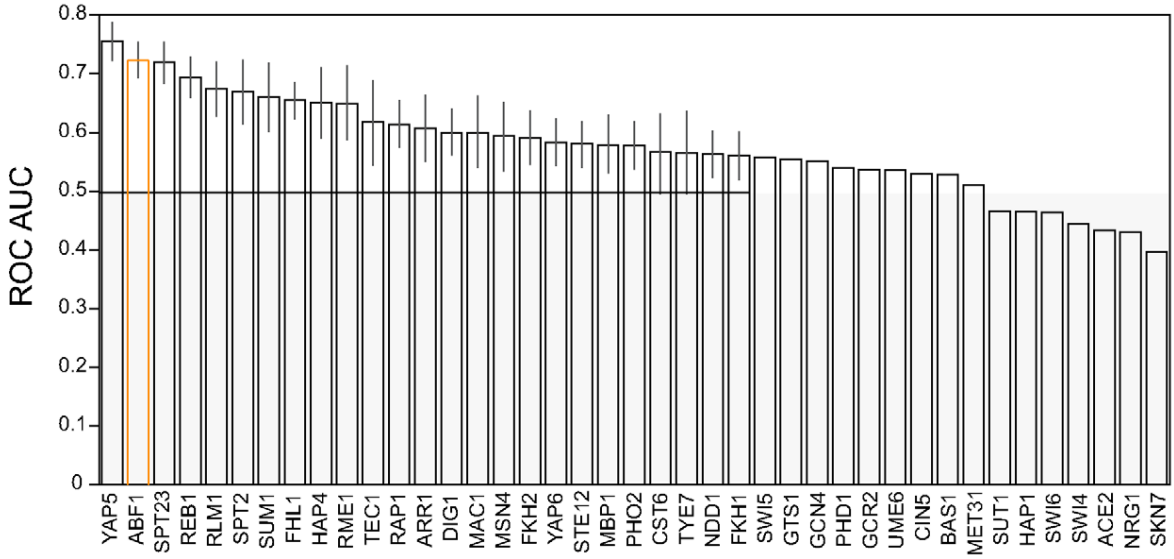


Рис. 4.3. Точность предсказания сайтов связывания транскрипционных факторов по площади под кривой ошибок ROC–AUC в геноме дрожжей на основе профиля плотности расположения нуклеосом по данным секвенирования [51].

Из рисунка 4.3 видно, что показатели точности не так высоки – около 0.7 в лучшем случае, а для некоторых сайтов даже меньше 0.5 – уровня, ожидаемого по случайным причинам. Распознавание Abf1 выделено на рисунке желтым цветом. Тем не менее, в целом присутствует существенная ассоциация между связыванием ТФ и плотностью нуклеосом в районе сайта.

Для рисунка 4.3 использовались данные, полученные без предварительной фиксации (перекрестного соединения) хроматина, но расчеты были выполнены также по данным экспериментов с фиксаций хроматина. Вопреки ожиданию, использование таких данных с фиксацией хроматина не улучшило результатов по ассоциации связывания ТФ и плотности нуклеосом. Чтобы исследовать этот эффект более детально, были построены распределения плотности нуклеосом на сайтах Abf1, полученные секвенированием нуклеосомной ДНК с предварительной фиксацией

хроматина и без таковой. Распределения хорошо согласуются друг с другом, выделяя локальные позиции соседних нуклеосом вокруг сайта связывания Abf1. Тем не менее, в районе самого сайта Abf1 наблюдается изменение – уменьшение плотности нуклеосом для нефиксированного хроматина (рис. 4.4).

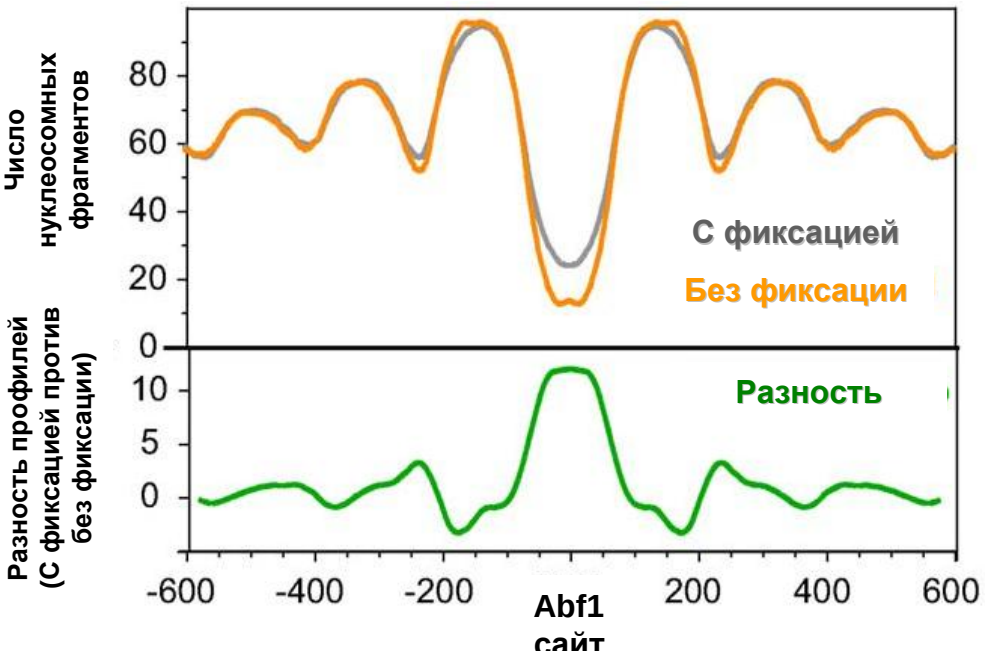


Рис. 4.4. Профили плотности нуклеосом в районе сайта Abf1 в геноме дрожжей, полученные секвенированием нуклеосомной ДНК с предварительной фиксацией (серый цвет) и без фиксации (желтый цвет) и карта разности профилей.

Полученный профиль разности между плотностью нуклеосом в двух экспериментальных состояниях может использоваться для предсказания самостоятельно.

Для более детальной оценки возможности предсказания сайтов связывания было выполнено сравнение предсказания, используя плотность нуклеосом с предварительной фиксацией и без фиксации хроматина. Действительно, избыток (разница) нуклеосомных фрагментов в профиле немного лучше предсказывает связывание ТФ с ДНК, чем только профиль плотности нуклеосом. 26 из 41 выборки ССТФ имели лучшие показатели распознавания точности под кривой ROC AUC, на основе разницы профилей. Этот эффект может быть связан с фиксацией нуклеосом, которые охватывают сайт связывания. Фиксация закрепляет нуклеосомы, скоротечно связывающиеся с конкретным участком ДНК, из-за большого числа аминов (лизинов и аргининов) лежащих вблизи ДНК. Такая фиксация нуклеосом в расположения сайта связывания ТФ может быть проиллюстрирована появлением очень сильного пика в

карте разности профилей нуклеосом, лежащего в район сайтов связывания Gal4 в промоторе GAL1–GAL10 (Рисунок 4.5).

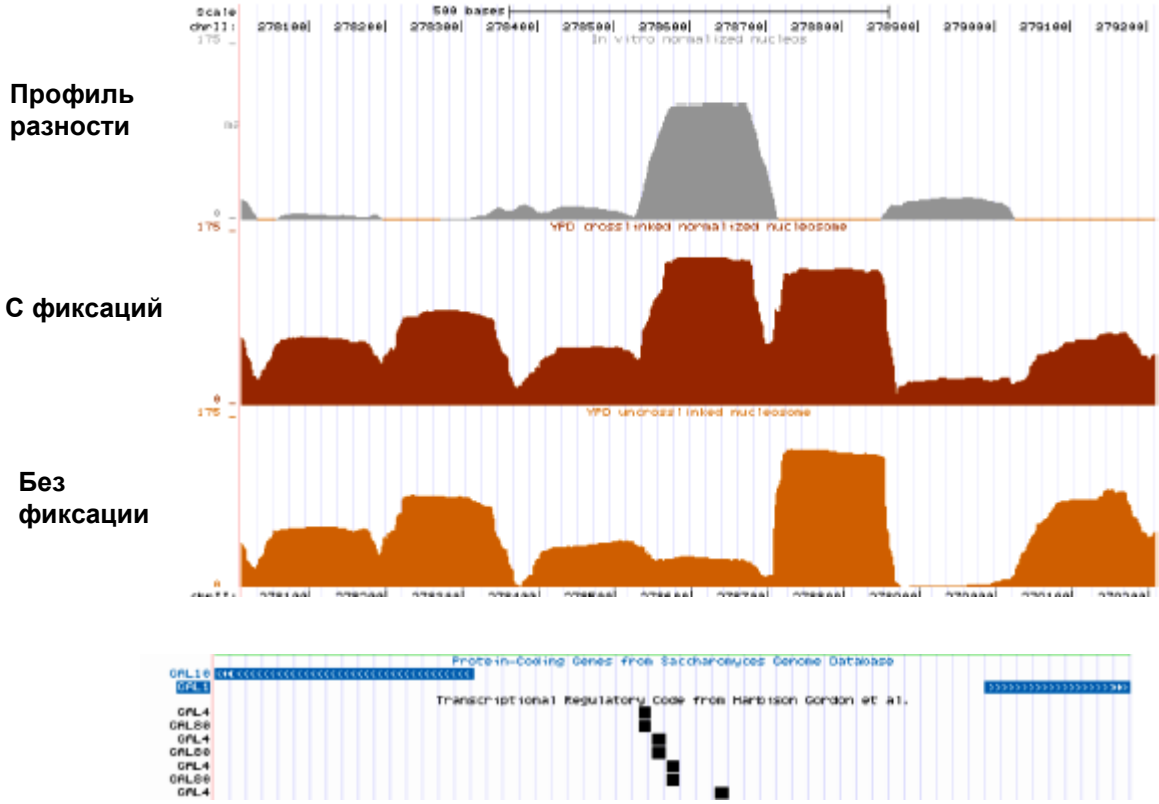


Рис. 4.5. Профили плотности нуклеосом в районе сайтов связывания Gal4 в промоторе GAL1–GAL10 – полученные с фиксацией и без фиксации хроматина (коричневый и желтый цвет, соответственно). Верхний профиль (серый цвет) представляет разность нуклеосомных профилей.

Из рисунка 4.5 видно, что нуклеосома присутствует в эксперименте с предварительной фиксаций (коричневый профиль в середине), и практически отсутствует в противном случае (нижний профиль), в то же время шесть соседних нуклеосом в данном районе имеют практически неизменный профиль плотности. В целом рисунок показывает характерный профиль расположения нуклеосом, в виде плавных широких волн, размером около 150 нуклеотидов, которые видно даже без дополнительной обработки профиля. Таким образом, независимо от механизма, ассоциация разность профилей секвенирования нуклеосом с предварительной фиксацией хроматина и без таковой имеет предсказательный эффект для определения связывания транскрипционный факторов, так же как и низкая плотность нуклеосом сама по себе.

Эффект контекстных свойств формирования нуклеосом на присутствие ТФ в консенсусных сайтах связывания, оцененный с помощью ChIP-qPCR.

В работе Kaplan et al. было сделано важное наблюдение, что геномные локусы, связанные транскрипционными факторами in vivo, имеют тенденцию быть свободными

от нуклеосом в восстановленном хроматине in vitro [264]. Таким образом, низкая плотность нуклеосом, наблюдаемая на сайтах связывания ТФ, присуща свойствам самих нуклеотидных последовательностей, контекстным свойствам ДНК, а не является только следствием конкуренции белковых факторов за связывание.

Для нескольких транскрипционных факторов способность их связывания их была выборочно проверена с помощью отдельных экспериментов ChIP-qPCR. Рисунок 4.6 показывает связь плотности расположения нуклеосом по данным секвенирования и силы связывания по ChIP-qPCR для ССТФ Dig1, Bas1, Gcn4, Nrg1 в геноме дрожжей в условиях in vivo и in vitro.

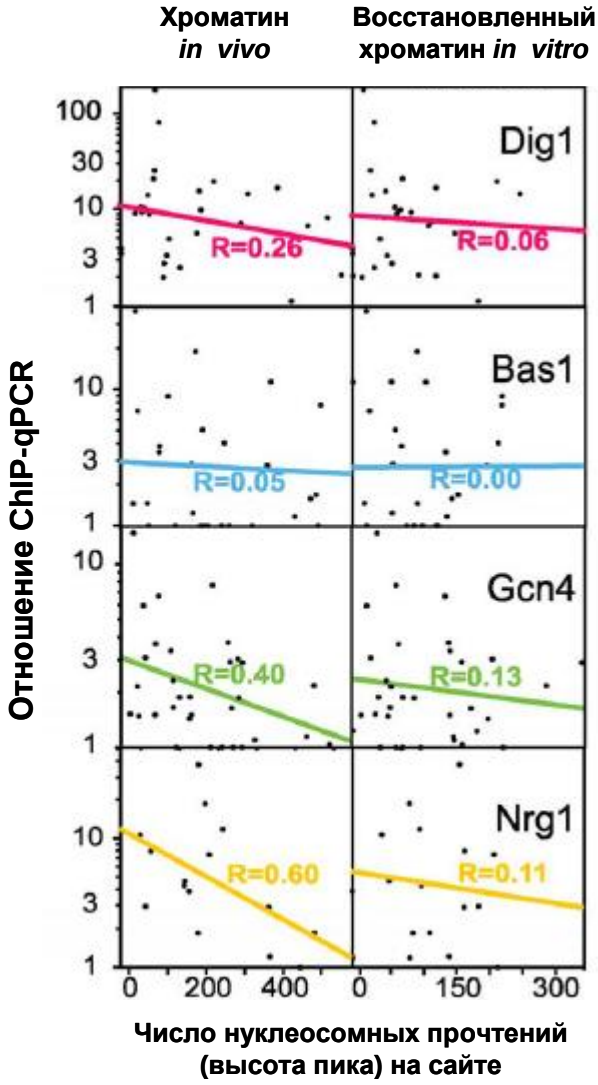


Рис. 4.6. Связь плотности расположения нуклеосом в ССТФ и силы связывания по ChIP-qPCR для факторов Dig1, Bas1, Gcn4, Nrg1 в геноме дрожжей в условиях in vivo и in vitro.

Из рисунка видно, что корреляция в условиях in vivo (левая панель) выше, чем in vitro (правая панель).

Был выполнен расчет предсказания сайтов по числу фрагментов нуклеосомной ДНК в профиле, используя оценку площади под кривой ошибок (ROC AUC). Для предсказания связывания ТФ, используя профили нуклеосом in vivo, точность предсказания выше, чем по профилям нуклеосом в in vitro восстановленном хроматине. Для 33 из 41 транскрипционного фактора, включая Abf1 и Reb1. Лучшая точность предсказания связывания in vivo показывает, что понижение плотности нуклеосом является следствием связывания транскрипционных факторов. В целом корреляция между значениями точности предсказания in vivo and in vitro достаточно высока (коэффициент корреляции R=0.79).

Было оценено связывание для выборок, содержащих от 16 до 33 консенсусных сайтов для каждого из четырех ТФ (Dig1, Bas1, Gcn4 и Nrg1). Действительно, повышенное ChIP связывание положительно коррелирует с низкой плотностью нуклеосом. Коэффициенты корреляции варьировали от 0.05 для Bas1 до 0.60 для Nrg1. Коэффициенты статистически значимы, исключая Bas1.

Были проверены также эффекты усреднения плотности нуклеосом в окне, варьирующем до 600 нт. Показано, что точность распознавания оптимальна до размера окна в 300 нт, затем значительно уменьшается.

В целом, показана возможность распознавания сайтов связывания по данным только плотности нуклеосомной упаковки, показано, что точность в целом невелика – около 0.6. Выявлен эффект лучшего предсказания связывания ТФ по данным in vivo, чем in vitro, и чем теоретически предсказанным в [51].

4.2. Исследование позиционирования нуклеосом и эффективности трансляции генов у дрожжей

Было проведено компьютерное исследование позиционирования нуклеосом и эффективности трансляции генов у дрожжей [63]. Для оценки неравномерности в использовании кодонов были предложены различные индексы, основанные на подсчете частоты кодонов, встречающихся в кодирующих частях генов, описанной ранее [63]. В *E.coli* и *S.cerevisiae* выбор синонимичного кодона в высокоэкспрессирующихся генах находится под строгим контролем естественного отбора, т.е., зная паттерн кодонов в этих генах, можно определить, какой из синонимичных кодонов наиболее эффективен для трансляции.

На уровне транскрипции эффективность экспрессии генов зависит, в том числе, от 5’- регуляторной области, в частности от локализации нуклеосом в ней. Контекстно-

направленное позиционирование нуклеосомы, обеспечиваемое правильными взаимодействиями функциональных сайтов ДНК с негистоновыми белками, играет важную функциональную роль среди факторов, определяющих регулярность формирования нуклеосом.

Для оценки распространенности механизмов оптимизации экспрессии генов в виде согласованности процессов на разных уровнях экспрессии исследовались корреляции плотности нуклеосомной упаковки в позициях 5'– нетранслируемых областей генов дрожжей видов *S. cerevisiae* и *S. pombe* со значением индекса эффективности элонгации соответствующих генов. Проверяемая гипотеза заключалась в следующем: для эффективной экспрессии генов необходимы согласованно оптимизированные процессы трансляции и транскрипции, в частности – инициации транскрипции и элонгации трансляции. Такая корреляция была найдена между ФНП в 5' нетранслируемых областях окрестности AUG кодона генов дрожжей видов *S. cerevisiae* и *S. pombe* со значением ИЭЭ соответствующих генов [62].

Для *S. cerevisiae* и *S. pombe* последовательности 5'- нетранслируемых областей окрестности старт-кодона трансляции генов от –600 до +600 нуклеотида относительно AUG кодона экстрагировались из генных карт этих организмов, взятых из базы данных GenBank (http://www.ncbi.nlm.nih.gov/). Экстракция проводилась на основе первичной разметки, имеющейся в этих генных картах с вырезанием интронов. Не рассматривались гены с неопределенными нуклеотидами (N) или длиной менее 30 кодонов. Также не рассматривались последовательности, маркированные как псевдогены. Анализ проводился для 5698 генов *S. cerevisiae* и 5088 генов *S. pombe*, а также для 15% генов этих выборок с максимальным и минимальным значениями индекса эффективности элонгации. Для *S.cerevisiae* использовались экспериментальные данные по полногеномному секвенированию нуклеосомной ДНК [264] в различных условиях. Использовались данные только для 5418 генов. Профиль плотности расположения нуклеосом в референсном геноме дрожжей был построен из коротких секвенированных фрагментов («ридов»). Данные секвенирования представлены автором в GEO NCBI (GSM351492). Только однозначно картируемые на геном дрожжей фрагменты, полученные с помощью технологии Illumina (Genome Analyzer II), были использованы при построении профиля плотности расположения нуклеосом. Для корректного компьютерного построения профиля каждый короткий фрагмент («рид») при картировании на хромосомы был увеличен до средней длины фрагмента, полученного при дроблении геномной ДНК. Чтобы построить профиль плотности расположения нуклеосом использовались исходные данные секвенирования для

картирования нуклеосомной ДНК, полученной *in vivo* при оптимальных условиях роста дрожжей (YPD). В итоге каждой нуклеотидной позиции в каждой хромосоме получила в соответствие значение плотности расположения нуклеосом – целое число соответствующее числу перекрывающихся фрагментов секвенирования, перекрывающих данную позицию.

Для двух исследуемых видов дрожжей был рассчитан индекс эффективности элонгации (ИЭЭ) с помощью программы EEI-Calculator (http://wwwmgs.bionet.nsc.ru/cgi-bin/mgs/eei-calculator/index.pl), разработанной в ИЦиГ СО РАН. Рассчитывается ИЭЭ по трем факторам: кодонному составу кодирующей последовательности, ее насыщенности инвертированными повторами, свободной энергии потенциальных шпилек. Для каждого организма находится оптимальная индивидуальная комбинации этих факторов.

Ранее были получены численные характеристики элонгации трансляции и распределение 759 одноклеточных организмов (702 бактерий, 52 архей, 5 эукариот) по классам. Показаны достоверные корреляции между значениями ИЭЭ и уровнем экспрессии по данным микрочиповых экспериментов в *S. cerevisiae* и *H. pylori*. Выходной файл этой программы содержит информацию о значениях индекса эффективности элонгации всех генов исследуемого организма, расположенных в порядке убывания. Для работы использовалась форма индекса ИЭЭ адекватно распознающая рибосомные гены как высокоэкспрессирующиеся [62].

Для 5’-района каждого гена был рассчитан нуклеосомный потенциал – функция, которая характеризует вероятность расположения нуклеосомы в заданном сайте последовательности. Значения этой функции вычисляются на основе частот динуклеотидов. Для расчета нуклеосомного потенциала использовалась программа RECON (http://wwwmgs.bionet.nsc.ru/mgs/programs/recon) [256]. Эта программа рассчитывает потенциал формирования нуклеосом (ПФН) в окне $W = 140$ нт по всей последовательности длиной L. Значение потенциала приписывается позиции центра окна. При анализе одной последовательности получается профиль нуклеосомного потенциала длины $L - W + 1$. Значения ПФН всегда меньше 1, а значение +1 соответствует наибольшей вероятности сайта формирования нуклеосомы.

Для каждого гена был получен профиль из 1061 значения ПФН для 1061 позиции. Далее вычислялся коэффициент корреляции между двумя векторами: вектор значений ПФН для определенной позиции фазированных относительно старта трансляции (AUG) всех последовательностей и вектор значений ИЭЭ для всех последовательностей. Для полученных результатов были рассчитаны критические

значения коэффициента корреляции при уровне значимости 0.95 с использованием критерия Фишера. Эти профили усредняли: для всех последовательностей, упорядоченных по значению ИЭЭ, для 15% имеющих максимальные значения ИЭЭ и для 15% имеющих минимальные значения.

Для *S. cerevisiae* коэффициент корреляции между ИЭЭ и экспериментально определенной эффективностью экспрессии составляет 0.8 при $p < 0.00001$. *Для S. cerevisiae* и *S. pombe* были рассчитаны коэффициенты корреляции между двумя векторами: вектор значений ПФН для определенной позиции фазированных относительно старта трансляции (AUG) всех последовательностей и вектор значений ИЭЭ для всех последовательностей. Профили коэффициентов корреляции между ПФН и ИЭЭ были получены для (–600; +600) участков, относительно старта трансляции экстрагированных из GenBank последовательностей генов.

Показана достоверная положительная корреляция между ИЭЭ и НФП для низкоэкспрессирующихся последовательностей в интервале (–250, –120), а также в (–590, –550) и в (–440, –390) относительно начала гена. И достоверная, и отрицательная корреляция – для высокоэкспрессирующихся последовательностей в интервале (–550, –520). Т. е., чем ниже эффективность элонгации кодирующей части, тем прочнее нуклеосомная упаковка в 5’-области. Высоко достоверная положительная корреляция между ИЭЭ и НФП для низкоэкспрессирующихся последовательностей в начале кодирующей части (0, +700), видимо, объясняется тем, что не только инициация, но и элонгация транскрипции для низкоэкспрессирующихся дрожжей тормозиться с помощью прочной нуклеосомной упаковки. Хотя, возможно, это следствие отбора кодонов в кодирующей части.

Была проверена корреляция ИЭЭ генов *S. cerevisiae* с экспериментальными данными по расположению нуклеосом в геномной ДНК, рассмотренными в предыдущем разделе, и в работе [264]. Был выполнен пересчет профиля плотности нуклеосом из [264], отфильтрованы полностью идентичные фрагменты и фрагменты, не имеющие однозначной локализации в геноме. Удлиненные до 146 нт фрагменты, что соответствует размеру нуклеосомы, формируют наложенный друг на друга профиль в хромосомных координатах, который и является плотностью, или картой расположения нуклеосом. Таким образом, каждая позиция на хромосоме имеет в соответствии число фрагментов секвенированной нуклеосомной ДНК, перекрывающих данную позицию. Это число экспериментальной плотности расположения нуклеосом может быть сравнено с любым другим числом, характеризующим данную позицию, например, на основе анализа окружающего контекста, либо также определенным в геноме экспериментально.

Корреляция между ИЭЭ и экспериментальными данными по нуклеосомной упаковке для *S. cerevisiae*. Были рассчитаны коэффициенты линейной корреляции

между ИЭЭ и экспериментальными данными по плотности нуклеосомной упаковки, полученной прямым секвенированием нуклеосомной ДНК для *S.cerevisiae* в районе промоторов генов, построен усредненный профиль. Для низкоэкспрессирующихся последовательностей в интервале (–110, 0) корреляция (коэффициент корреляции до 0.25) достоверна и положительна. Для высокоэкспрессирующихся последовательностей корреляция (коэффициент корреляции до –0.15) в этом же интервале достоверна и отрицательна.

Эти данные подтверждают проверяемую гипотезу, хотя по локализации областей значимых корреляций не вполне совпадают с найденными, исходя из теоретических оценок. Отличие нуклеосомной локализации определенной по экспериментальным данным *in vivo* от теоретически предсказанных в промоторах, особенно в районе (–200;0) было показано в предыдущем разделе и в работе [51].

Было установлено существенное повышение числа политрактов по сравнению с более дистальным районом промотора (–500;–250) (поли(А)$_5$). Трактов подобного типа практически не наблюдается в районе ниже старта трансляции (0;+600). Также было обнаружено, что плотность поли(А)-трактов в промоторном районе всех генов заметно превышает плотность подобных трактов в последовательностях сайтов формирования нуклеосом дрожжей, а также в случайных последовательностях (марковские цепи нулевого порядка). Эти случайные последовательности были сгенерированы на основе нуклеотидного состава как известных сайтов формирования нуклеосом дрожжей, так и исследованных проксимальных промоторных районов.

Для промоторов *S. cerevisiae* между ИЭЭ и наличием поли(А)-трактов в последовательностях показана достоверная положительная корреляция, в частности для тракта поли(А)$_5$ в районе (–250; –50) коэффициент корреляции $r = 0.065$ ($p < 10^{-7}$). Аналогичный коэффициент для *S. pombe* статистически недостоверен [62].

Компьютерный анализ показал, что по сравнению с ожиданием на основе динуклеотидного состава последовательностей для вида *S. cerevisiae* наблюдается умеренное значимое обогащение трактов в дистальном районе (–600;–500), умеренное обеднение трактов поли(А)$_5$ в районе ниже старта трансляции (+100;+200) и значительное обогащение в проксимальном районе выше старта (–200;+100). В аналогичных районах генов *S. pombe* значимых обогащения и обеднения не наблюдается. Ранее для различных модельных организмов эукариот была показана связь между короткими олигонуклеотидами и экспериментально определенным положением нуклеосом [263], в частности, негативная корреляция нуклеосом с содержанием поли(А)-трактов. Это подтверждает найденную независимо корреляционную зависимость между ИЭЭ и поли(А)-трактами. Таким образом, подтверждена взаимосвязь между локализацией нуклеосом в

промоторных районах и индексом эффективности элонгации [62]. Индекс эффективности элонгации отрицательно коррелирует с потенциалом формирования нуклеосом.

В целом, проведенное исследование подтверждает положительную ассоциацию между открытым состоянием хроматина в промоторе и эффективностью транскрипции гена, оцененной с помощью рассмотренных теоретических индексов.

4.2. Исследование ассоциации сайтов связывания ТФ с модификациями хроматина

Существенно новым шагом геномной составляющей предсказания активных сайтов связывания транскрипционных факторов (ТФ) в геноме человека являются данные о геномном «ландшафте» хроматина, включая модификации гистонов, экспериментальное исследование которых возможно с помощью высокопроизводительного секвенирования ДНК и иммунопреципитации хроматина - ChIP-seq. Паттерны такого ландшафта в геноме человека в различных тканях, в первую очередь в раковых клетках, существенно различны.

С помощью компьютерной программы поиска совпадений было показано, что из огромного набора потенциальных сайтов связывания для ТФ рецептора эстрогенов ERα в геноме человека, используя определение только по последовательности 13 нуклеотидов, можно получить около миллиона сайтов. Лишь малая часть сайтов связана in vivo (1-2%, около 17 тысяч сайтов) [13]. Более точно, для рецептора эстрогенов число сайтов связывания в геноме человека составляет около 1 миллиона по консенсусу (с несовпадениями), около 32 тыс. сайтов по позиционной весовой матрице, и число пиков, экспериментально определенных сайтов с помощью ChIP-seq около 17 тысяч (в культурах раковых клеток MCF-7 и T47D). Вопрос состоит в выявлении причин такого относительно малого числа «активных», или реально связанных сайтов в геноме. Общий ответ здесь не в контекстных характеристиках, кодирующих сайты, а в эпигенетических модификациях, геномном «ландшафте», определяющем доступность ДНК для связывания с белками. Более точно формировать доступность сайтов для связывания могут и другие белковые факторы, или ко-факторы, связывающиеся с ДНК и облегчающие доступ транскрипционного фактора.

С помощью перебора и комбинации 12 генетических и эпигенетических параметров было найдено, что специфичность связывания ERα в геноме обеспечивают мотив элемента ответа на эстроген (ERE - estrogen response element), присутствие сайта связывания транскрипционного фактора FOXA1, присутствие маркера модификации гистонов H3K4me1 и открытая конфигурация хроматина. Эти факторы могут

предсказать вызванное эстрогеном связывание ERα с высокой аккуратностью: значения площади под кривой ROC-AUC составляют 0.95 и 0.88 при использовании таких геномных характеристик для построения решающих правил предсказания [13]. Более того, при оценке предсказания связывания рецептора эстрогенов в другой эстроген-позитивной клеточной линии, та же модель показала высокую точность предсказания связывания ERα (ROC-AUC составило 0.86). Вариабельность в выборе сайта связывания между клеточными линиями MCF-7 и T47D определяется сайтами с субоптимальным мотивом ERE, и модулируется состоянием хроматина в клетках [13]. Эти результаты предполагают определенные связи между мотивами в последовательности ДНК и локальной конфигурацией хроматина в составе нуклеосом при выборе связывания транскрипционного фактора.

Для исследования использовались данные о структуре хроматина - модификации гистона H3, полученные с помощью ChIP-seq. Такие данные доступны в GEO NCBI для следующих модификаций гистона 3 - моно- и триметилирование лизина в позициях 4, 9, 27, 36, ацетилирование в позициях 9, 14 (аббревиатура модификаций представлена как – H3K4me3, H3K4me1, H3K27me3, H3K9me3, H3K9ac, H3K14ac). Ранее в работе автора было показано, что метилирование гистонов в позиции 4 [19], ацетилирование в позиции 9 и 14 являются индикаторами активной транскрипции, и показательны для сайтов связывания активирующих транскрипцию генов в промоторных участках, в которых они находятся. В то же время триметилирование лизина в позициях 9 и 27 гистона 3 (H3K27me3, H3K9me3) указывает на подавление экспрессии близлежащих генов. Модификации хроматина показывают состояние нуклеосомной упаковки хромосомных регионов, соответственно, степень доступности таких участков для связывания факторов транскрипции и комплекса РНК полимеразы II. Прямое секвенирование нуклеосомных последовательностей ДНК также становится технически доступно для генома человека (клетки CD4 и отдельные хромосомы). Новые данные становятся доступны в рамках международных проектов ENCODE [8], EpiGRAPH. Эпигенетические модификации хроматина, передаваемые в последовательности клеточных делений, могут быть сопоставлены с результатами определения сайтов связывания транскрипционных факторов по данным ChIP-seq.

Данные по размерам библиотек секвенирования ChIP-seq для определения сайтов модификаций гистонов и ССТФ в клеточных линиях MCF-7 и T47D [13] представлены в таблице.

Таблица 4.1

Размер библиотек ChIP-seq для транскрипционных факторов и модификаций хроматина в клеточных линиях MCF-7 и T47D

Транскрипцион-ный фактор/ Модификация хроматина	**Клеточная линия**	**Обработка клеток**	**Антитело**	**Размер библиотеки (в тысячах прочтений)**
ER α	MCF-7	E2	Cat# HC-20, Santa Cruz	7,009
		растворитель		12,658
	T47D	E2		7,640
		растворитель		11,724
FAIRE	MCF-7	E2	-	12,612
		растворитель		12,287
	T47D	E2		20,258
		растворитель		14,860
RNA Pol II	MCF-7	E2	Cat# ab5408, Abcam	7,556
		растворитель		9,556
H3K9me3	MCF-7	E2	Cat# ab8898, Abcam	13,789
		растворитель		14,846
H3K27me3	MCF-7	E2	Cat# 07-449, Upstate Biotechnology Inc.	17,253
		растворитель		14,686
H3K4me1	MCF-7	E2	Cat# ab8895, Abcam	7,705
		растворитель		10,171
	T47D	E2		18,377
		растворитель		17,067
H3K4me3	MCF-7	E2	Cat# ab8580, Abcam	16,962
		растворитель		14,162
H3K9ac	MCF-7	E2	Cat# 07-352, Upstate Biotechnology Inc.	8,527
		растворитель		7,600
H3K14ac	MCF-7	E2	Cat# 07-353, Upstate Biotechnology Inc.	11,174
		растворитель		9,276
FOXA1	MCF-7	E2	Cat# AB4124, Chemicon	13,182
		растворитель		17,932
	T47D	E2		6,764
		растворитель		14,860
c-Fos	MCF-7	E2	Cat# sc-7202, Santa Cruz	18,222
		растворитель		15,261
c-Jun	MCF-7	E2	Cat# sc-45, Santa Cruz	15,696
		растворитель		14,632

Таблица 4.1 показывает, что размеры библиотек секвенирования варьировали от 7 до 20 миллионов прочтений. Все эксперименты были сделаны в парах: клетки линии MCF-7 до активации (в растворителе DMSO) и после активации эстрадиолом (E2). Часть экспериментов была сделана также в клетках линии T47D [13].

Кроме того, для анализа доступности ДНК (плотности нуклеосомной упаковки) в сайтах связывания исследуемых ТФ, экспериментально была определена

конфигурация доступности хроматина в геноме с помощью секвенирования фрагментов геномной ДНК, изолированных по методу FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) [271], выделения регуляторных элементов с помощью формальдегида. Метод FAIRE включает обогащение последовательностей ДНК, свободных от нуклеосом, в водной фазе после экстракции фенолом [271]. Число таких выделенных и секвенированных FAIRE фрагментов ДНК для исследуемого участка генома соответствует степени освобождения данного участка от нуклеосом.

Используя сайты связывания ERα, выделенные в эксперименте ChIP-seq, были проанализированы характеристики конфигурации хроматина (полногеномные профили секвенирования участков модификаций гистонов) для локусов, содержащих сайты связывания ERα. Был выполнен компьютерный анализ присутствия в геноме ChIP-seq маркеров модификаций гистонов до и после обработки эстрогеном (E2): РНК-полимеразы II, маркеры активации транскрипции H3K4me1, H3K4me3, H3K9ac и H3K14ac, и маркеры репрессии транскрипции H3K9me3 и H3K27me3 (рисунок 4.7).

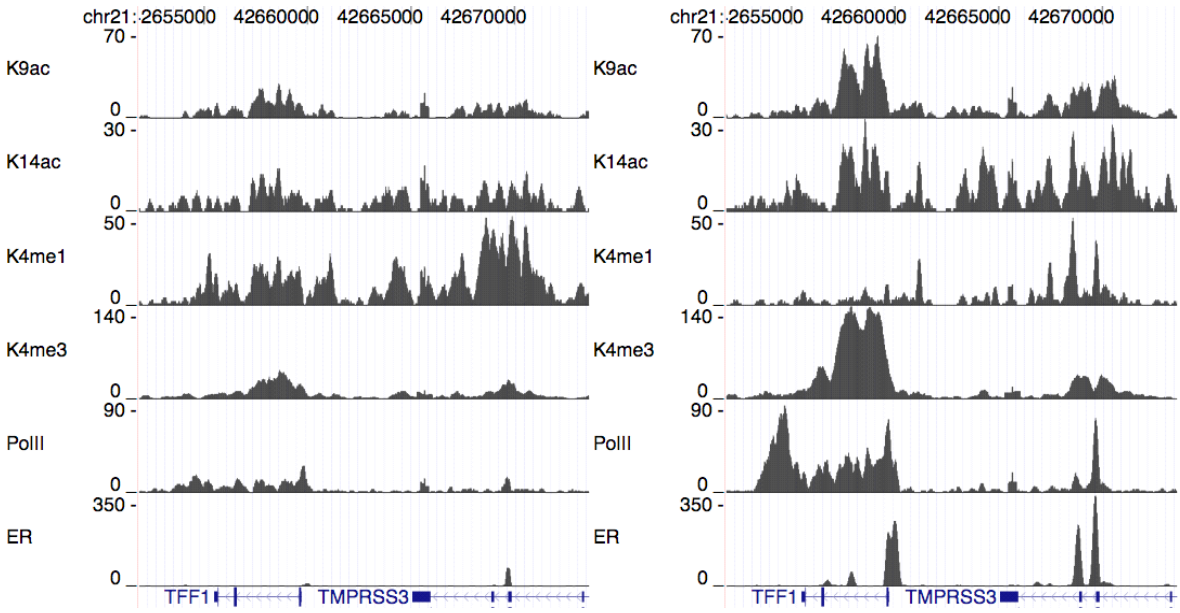


Рис. 4.7. Профили ассоциации маркеров открытого хроматина с сайтами связывания ERα в геноме человека для гена TFF1 до и после активации эстрогеном (обработки клеток E2).

Видно появление ChIP-seq пика связывания ERα в районе промотора гена TFF1 (правая панель). Из рисунка видно сопутствующее изменение геномного ландшафта модификаций (увеличение высоты профилей почти для всех модификаций на правой панели). Профиль секвенирования FAIRE соответствует участкам открытого хроматина. Действительно, анализ профилей FAIRE показал, что хроматин открыт в радиусе 1Кб вокруг сайтов связывания ERα, как до, так и после обработки эстрадиолом (E2).

На следующем рисунке 4.8 показаны усредненные профили маркеров шести модификаций хроматина ChIP-seq, а также FAIRE и полимеразы II, построенные для геномных окрестностей сайтов связывания ERα. Представлено два профиля для каждого маркера состояния хроматина - до и после активации эстрогеном экспрессии ERα в клетках MCF-7, всего 8 пар профилей.

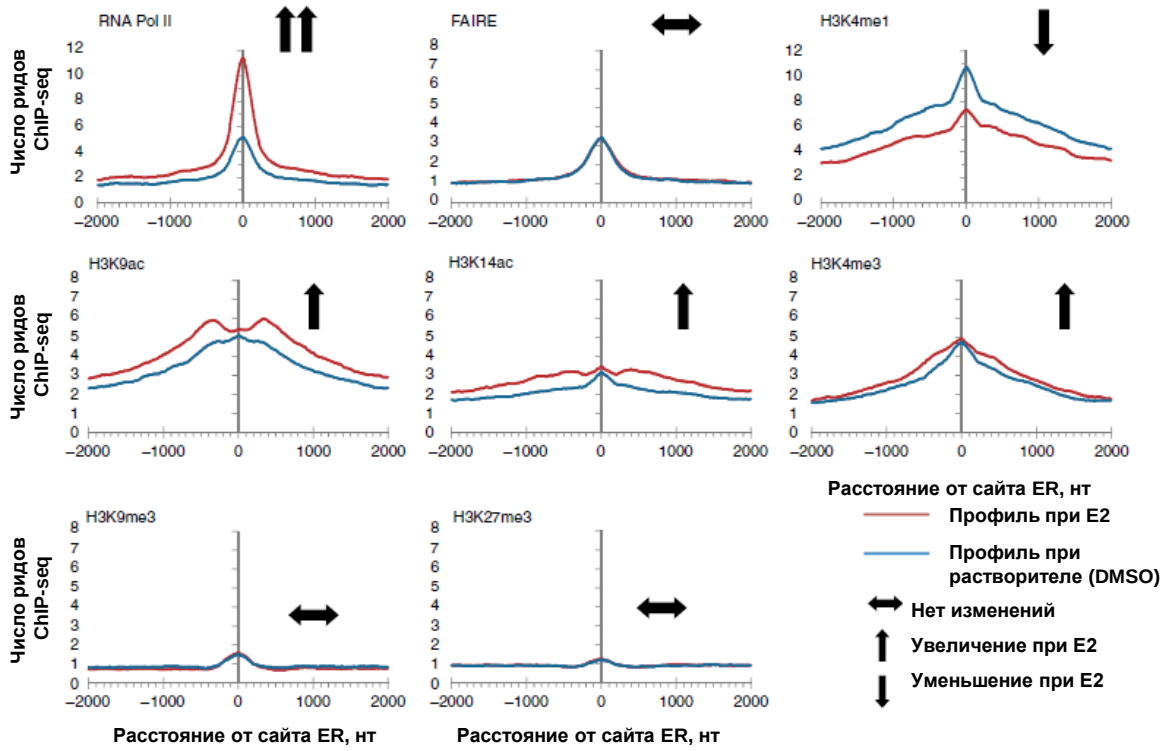


Рис. 4.8. Ассоциация модификаций хроматина с сайтами связывания ERα в геноме человека в клеточной линии MCF-7. Эстрадиол (E2) индуцирует изменения хроматина вокруг сайтов связывания ERα (красные линии против синих линий). Стрелками показан эффект изменения профиля ChIP-seq для модификаций гистонов для различных маркеров.

Из рисунка видно, что после обработки эстрадиолом (E2) вызывающей экспрессию ERα профили модификаций гистонов - маркеров активации - H3K4me1, H3K4me3, H3K9ac и H3K14ac вокруг сайтов связывания ERα увеличиваются (обозначено стрелками). Увеличивается также профиль РНК-полимеразы II. В то же время маркеры, связанные с репрессией транскрипции, H3K9me3 и H3K27me3 не меняются при обработке клеток эстрадиолом.

Прочтения ДНК из каждой библиотеки ChIP-seq (после нормализации до 7 миллионов прочтений для сопоставления) были использованы для расчета среднего числа прочтений на нуклеотид в интервале 2Кб относительно центра найденных ChIP-seq сайтов связывания ERα. Все 16,043 сайтов связывания ERα, определенные в пиках ChIP-seq были сортированы в убывающем порядке по размеру пика и подразделены на четыре равные группы - квартили (Q1–Q4). Группа первой квартили содержит наиболее

сильные сайты связывания ERα, в то время как 4-я квартиль содержит наиболее слабые сайты (цветные линии обозначают квартили: Q1, Q2, Q3 и Q4). Для расчетов профилей на рис. 4.8 использовались позиции на хромосоме сайтов связывания ERα из квартили 1 (из ранжированного списка по высоте пиков ChIP-seq, наиболее сильные сайты связывания).

Наблюдается также градиент изменения открытости хроматина при переходе от квартиля к квартилю для сайтов связывания ERα, при этом квартиль 1 (наибольшая сила связывания ERα) показывает максимальную открытость хроматина с пиком в центре сайта, а квартиль 4 (наименьшее связывание ERα) показывает наименьший профиль. На следующем рисунке показаны профили для маркеров активации состояния хроматина для сайтов связывания ERα после стимуляции эстрогеном (E2) по квартилям. Профили модификаций гистонов, фазированы на сайтах связывания ERα в геноме человека, определенных с помощью ChIP-seq (16 тысяч сайтов) и усреднены для каждой из четырех квартилей (рис. 4.10). Четыре цветные линии – сайты связывания ERα, разделенные по квартилям, ранжированные по высоте пика ChIP-seq. Серая линия – несвязанные с ERα участки ДНК.

Видно, что более сильные сайты связывания ERα имеют более высокий профиль маркеров открытого хроматина, т.е. наблюдается качественное соответствие между силой связывания транскрипционного фактора с ДНК по данным ChIP-seq и открытостью хроматина.

Из рисунка видно, что усредненные ChIP-seq профили модификаций гистонов (H3Kme1, H3K4me3, H3K9ac, H3K14ac) имеют более широкий пик связывания, порядка тысячи нуклеотидов вокруг сайтов связывания, что соответствует более широкому, чем участок связывания ТФ району содержащему несколько нуклеосом.

Присутствие этих гистоновых маркеров указывает точно на центр сайта связывания ERα и количественно коррелирует со связыванием ERα (рисунок 4.9).

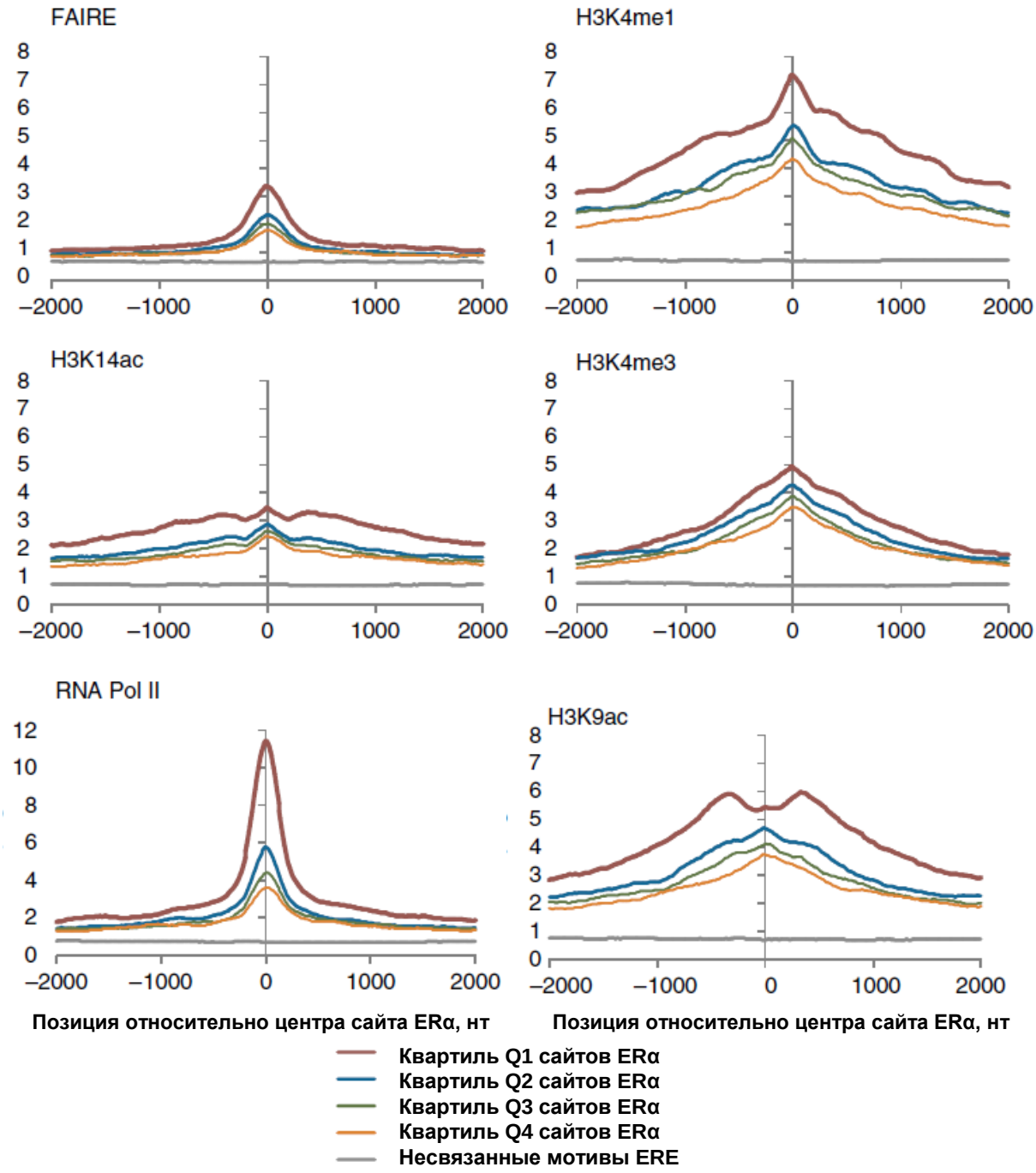


Рис. 4.9. Профили ChIP-seq модификаций хроматина, маркирующих активацию транскрипции, для сайтов связывания ERα в геноме человека, ранжированных по квартилям Q1-Q2-Q3-Q4. По оси ординат - высота пика ChIP-seq. Прочтения ДНК из библиотек ChIP-seq соответствующих активированному хроматину были использованы для расчета среднего числа прочтений на нуклеотид в интервале +/-2Кб относительно центра найденных ChIP-seq сайтов связывания ERα. Для сравнения, профиль для среднего числа прочтений на 10000 геномных сайтов, не связанных с ERα, показан серой линией.

Примечательно, что сигнал метилирования H3K4me1 в отсутствии эстрадиола также коррелирует со связыванием ERα, что соответствует представлениям об ассоциации H3K4me1 с участками с энхансерной функцией. Такая ассоциация была

отмечена в клетках HeLa человека для сайтов связывания p300, сайтов связывания STAT1, предсказанных энхансерах и сайтах FOXA2 из клеток печени мыши [316].

Похожим образом маркеры ацетилирования H3K9 и H3K14 значимо присутствуют в районах вокруг сайтов связывания ERα. Сигнал маркеров активации (открытого состояния хроматина) коррелирует с сигналом связывания ERα в сайтах связывания ERα. В противоположность с сайтами связывания ERα, выявленными в ChIP-seq эксперименте, никакие маркеры активации гистонов не давали сигнал для участков несвязанных сайтов, содержащих мотив ERE.

Из рисунка видно, что профиль метилирования и ацетилирования гистона H3, измеренные по маркерам H3K4me1, H3K4me3, H3K9ac, H3K14ac, имеют максимум в районе локализации сайтов связывания рецептора эстрогена ERα в геноме человека [13]. В данном случае речь идет о согласованности генетических сообщений, одно из которых (записанное в ДНК) определяет локализацию сайтов связывания ERα, а второе, записанное в структуре хроматина, определяет такой характер модификации гистонов, который обеспечивает его максимальную открытость в местах локализации ССТФ ERα.

Следующий рисунок 4.10 показывает профили ChIP-seq профили модификаций гистонов, относящихся к закрытому состоянию хроматина - H3K9me3, H3K27me3, для сайтов связывания ERα в тех же обозначениях, что и на предыдущем рисунке.

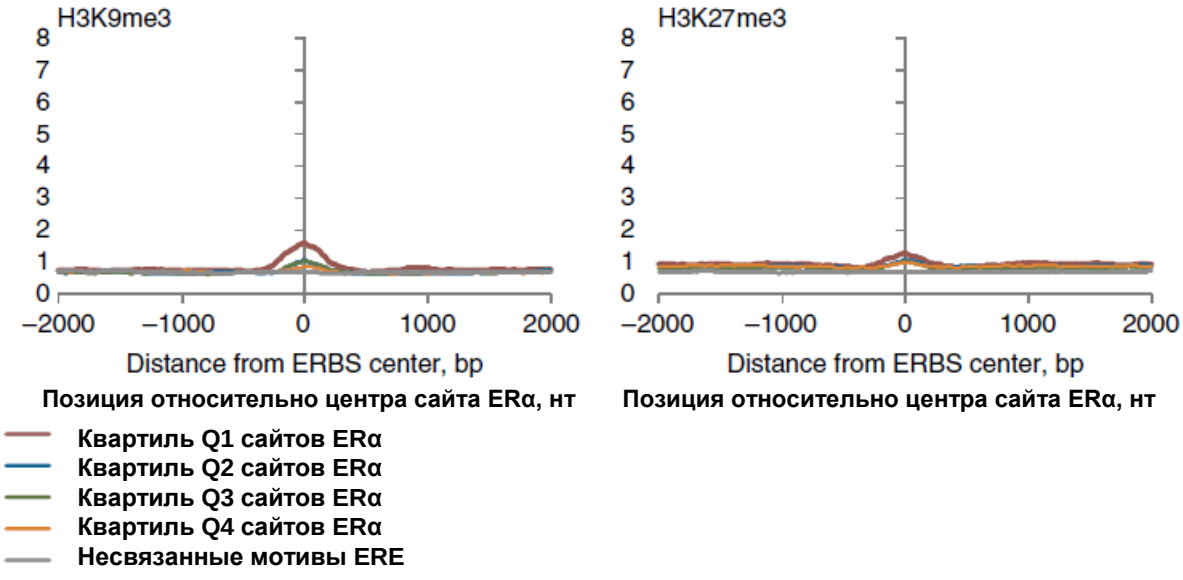


Рис. 4.10. Профили ChIP-seq модификаций хроматина, маркирующих закрытое состояние хроматина H3K9me3 и H3K27me3, для сайтов связывания ERα в геноме человека, ранжированных по квартилям. Обозначения соответствуют предыдущему рисунку.

Из рисунка видно отсутствие выраженного пика связывания, и близость профилей для всех 4 квартилей сайтов связывания. Таким образом, качественно

показано, что нет ассоциации сайтов связывания с маркерами репрессивного состояния хроматина.

Для исследования вопроса о том, насколько сильна ассоциация сайтов связывания ERα с маркерами модификаций гистонов в зависимости от расположения сайтов относительно генов, т.е. обогащены ли маркерами активного хроматина только промоторные или дистальные сайты, была выполнена следующая группировка сайтов. Вместо упорядочения (ранжировки) по квартилям относительно высоты пика все сайты ERα, выявленные в ChIP-seq эксперименте были классифицированы как промоторные сайты, внутригенные сайты, дистальные сайты, удаленные районы генома - «генная пустыня». Аналогично предыдущим рисункам были построены профили ChIP-seq маркеров хроматина для этих группам (рис. 4.11). Расстояние показано по оси X, а среднее число сайтов показано на оси Y для каждой панели.

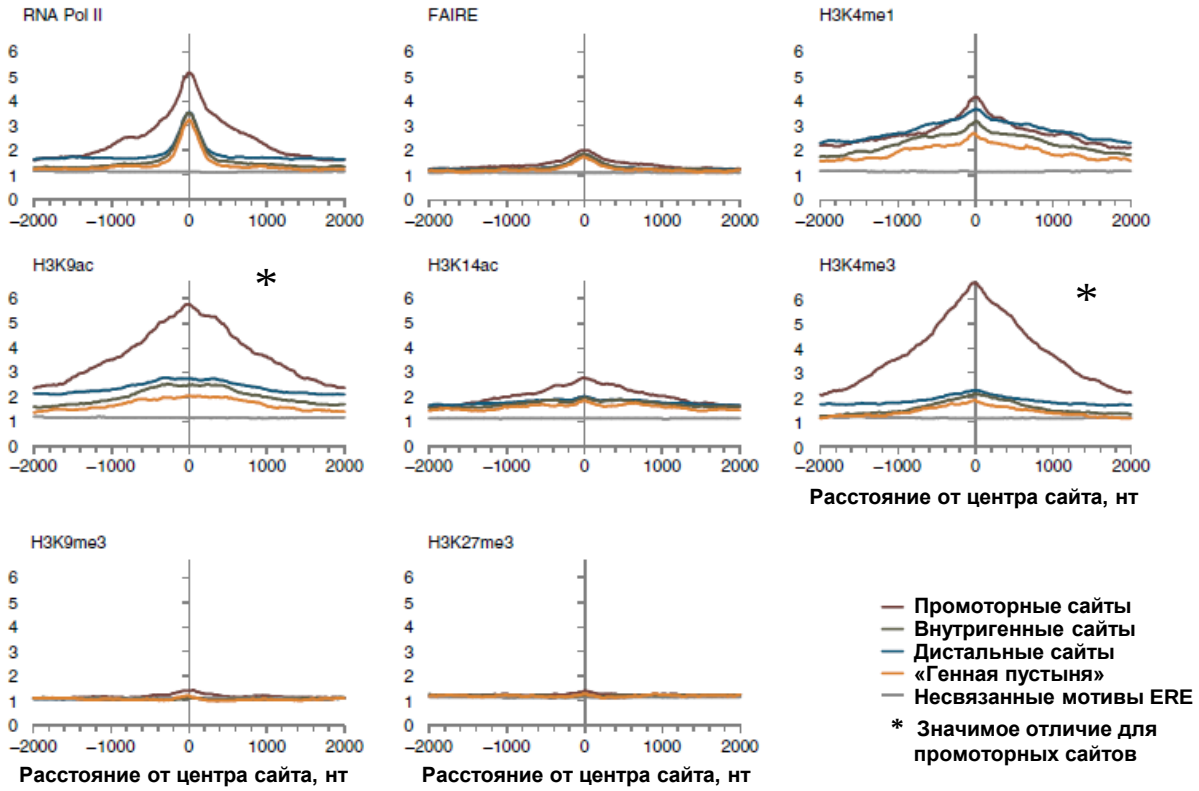


Рис. 4.11. Профили модификаций хроматина для проксимальных и дистальных сайтов связывания ERα в геноме человека (клетки MCF-7 после стимуляции E2). Представлены категории сайтов по расположению относительно генов - промоторные сайты, внутригенные сайты, дистальные сайты, удаленные районы генома - «генная пустыня». Профиль соответствует среднему значению профилей ChIP-seq в районе +/- 2Кб вокруг сайтов. Для сравнения приведен такой же средний профиль ChIP-seq для 10 000 случайных геномных сайтов, не связанных ERα (серые линии).

Использовалось следующее определение геномной локализации сайтов относительно генов RefSeq: промоторные сайты (5 Кб перед стартом транскрипции, внутригенные сайты (расположенные внутри границ гена), дистальные сайты (5–100 Кб

перед стартом транскрипции) и все остальные районы («генная пустыня»). Рисунок показывает, что промоторные сайты имеют наибольшее обогащение сигнала ChIP-seq для всех маркеров активации - РНК полимеразы II, FAIRE, H3K4me3, H3K4me1, H3K14ac, H3K9ac. В то же время другие сайты, расположенные дистально по отношению к генам, имеют повышенный средний сигнал для профилей активации в районе сайта, всегда выше, чем для сайтов не связанных с ERE (см. рисунок).

Таким образом, не только промоторные сайты, где можно ожидать присутствие маркеров активации хроматина, но и дистальные сайты вносят вклад в средний сигнал профиля. Открытая структура хроматина важна для связывания ERα с ДНК во всех районах относительно генов. Отметим, что маркеры модификаций гистонов H3K9me3 и H3K27me3, связанные с закрытым состоянием хроматина, не дают повышения профиля ChIP-seq ни для одной из категорий сайтов и неотличимы от уровня сигнала для случайных сайтов (см. нижние панели рисунка).

Рассмотрим ассоциацию сайтов связывания ERα с сайтами FoxA1, также определенных с помощью ChIP-seq в геноме. Известно, что FoxA1 содействует связыванию ERα, и можно ожидать, что более сильные ChIP-seq сайты ERα будут чаще содержать ChIP-seq сайты FoxA1. Рисунок 4.12 представляет долю (от 0 до 1) присутствия FoxA1 в сайтах ERα, разбитых по квартилям.

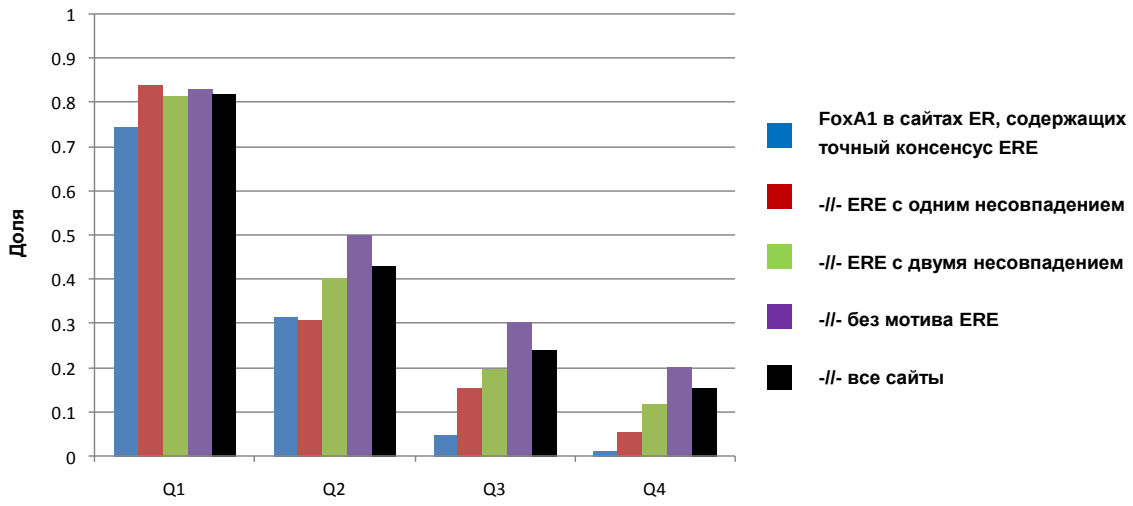


Рис. 4.12. Фракция содержания мотива FoxA1 в участках связывания ERα ранжированных по квартилям высоты пика ChIP-seq (от верхней по силе связывания квартили Q1 до Q4) в зависимости от силы мотива ERE (числа несовпадений с консенсусом).

Действительно, доля содержания FoxA1 достаточно высока (около 0.8) для первой квартили, понижаясь для других квартилей. В то же время наблюдается другой эффект - для сайтов связывания ERα с вырожденным мотивом ERE (с одним и двумя совпадениями) доля содержания FoxA1 возрастает, причем особенно заметно для

квартилей Q2, Q3, Q4. Таким образом, FoxA1 помогает связыванию в сайтах со слабым и более вырожденным мотивом ERE.

Статистическое подтверждение ассоциации между связыванием ТФ и изменением структуры хроматина (доступности) получено с помощью анализа маркеров активации и корреляции высоты ChIP-seq пика маркеров с высотой ChIP-seq пика ERα для всех 16 тысяч сайтов. Рассчитывались коэффициенты корреляции до активации эстрадиолом и после (E2). Коэффициент корреляции - общая мера связи присутствия маркеров открытого хроматина, численно представленная для каждого сайта числом прочтений ChIP-seq данного геномного маркера в геномной окрестности сайта, и высоты пика профиля связывания транскрипционного фактора ERα, также представленного численно числом прочтений. Использовался ранговый коэффициент корреляции Кендалла. Статистически значимая ассоциация сайтов связывания и активирующих модификаций хроматина показана на рисунке 4.13 и в таблице 4.2. Для маркеров репрессированного состояния хроматина H3K9me3 и H3K27me3, а также для контрольного секвенирования ДНК без иммунопреципитации, коэффициенты корреляции были близки к нулю и статистически не значимы.

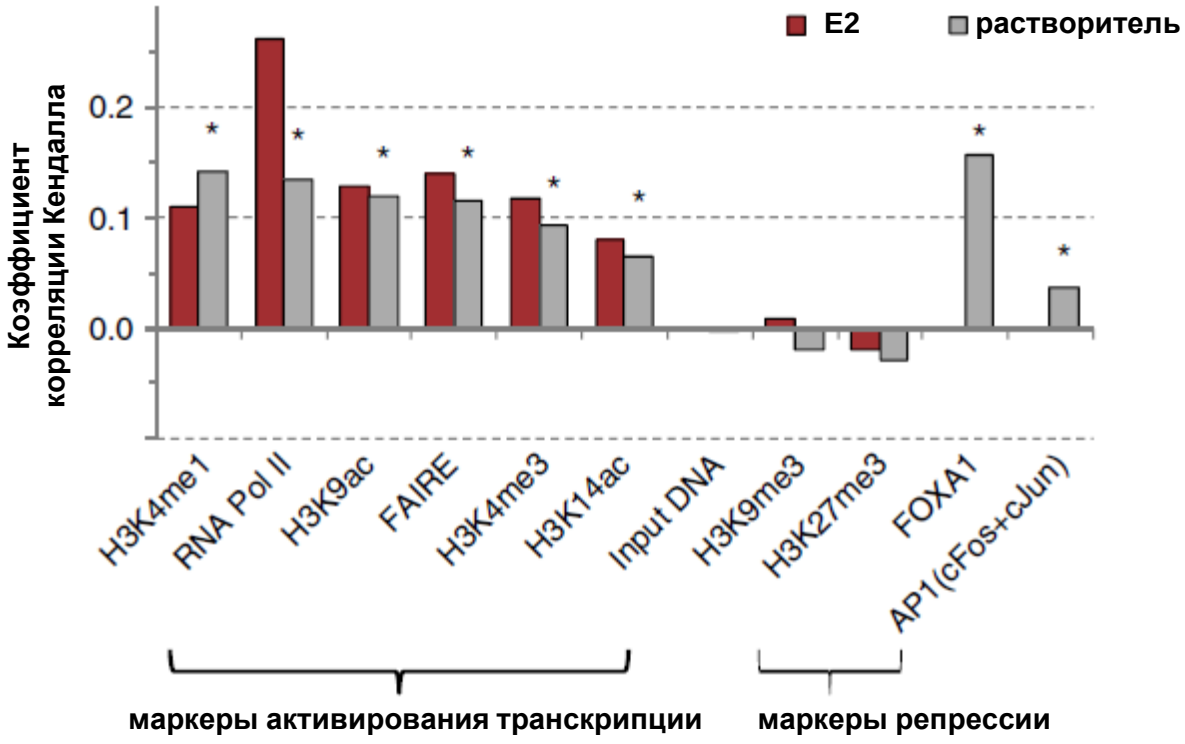


Рис. 4.13. Корреляция модификаций хроматина со связыванием ERα в геноме человека (корреляция интенсивностей связывания, оцененных по пикам ChIP-seq) [13].

Из рисунка 4.13 видно, что корреляция есть, она значима для обоих состояний. Такие оценки были использованы для предсказания сайтов связывания in vivo в геноме, с высокой точностью, как показано в работе [13].

Чтобы проверить, связан ли высокий коэффициент корреляции с размером библиотеки, и сравнивать такие величины между собой была выполнена процедура нормализации. Использовалось случайным образом выбранные (с помощью специально написанной для этого компьютерной программы, использующей датчик случайных чисел) 1 миллион прочтений, 5 миллионов прочтений и все прочтения библиотеки ChIP-seq (как правило, более 7 миллионов прочтений), затем коэффициент корреляции числа прочтений с высотой пика ChIP-seq для ERα был пересчитан.

Следующая Таблица 4.2 показывает величину корреляции и статистическую значимость для сравнения высоты ChIP-seq пиков ERα и числа прочтений ChIP-seq из библиотек модификаций хроматина.

Таблица 4.2

Коэффициент ранговой корреляции Кендалла и значимость для сравнения высоты ChIP-seq пиков ERα и числа прочтений библиотек модификаций хроматина, с учетом нормализации размера библиотек (1 миллион, 5 миллионов и полный размер)

Библиотека ChIP-seq	Коэффициент ранговой корреляции Кендалла (тау)			Уровень значимость*(P-value)*		
	1 млн. ридов	5 млн. ридов	Все риды	1 млн. ридов	5 млн. ридов	Все риды
H3K4me1	**0.1308**	**0.1416**	**0.1512**	**2.54E-136**	**2.12E-159**	**1.49E-181**
RNA Pol II	**0.095**	**0.1348**	**0.1452**	**8.26E-73**	**1.17E-144**	**1.62E-167**
H3K9ac	**0.0921**	**0.1196**	**0.1352**	**1.71E-68**	**3.20E-114**	**1.80E-145**
FAIRE	**0.0786**	**0.1163**	**0.1469**	**2.02E-50**	**4.16E-108**	**2.14E-171**
H3K4me3	**0.0639**	**0.093**	**0.1187**	**7.16E-34**	**6.90E-70**	**1.49E-112**
H3K14ac	**0.0389**	**0.0655**	**0.0762**	**1.49E-13**	**1.50E-35**	**5.05E-07**
Контрольное секвенирование	0.002	-0.0001	-0.0010	7.08E-01	9.82E-01	8.46E-01
H3K9me3	-0.0052	-0.0185	-0.0264	3.26E-01	**4.46E-04**	**5.05E-07**
H3K27me3	-0.0163	-0.0287	-0.0323	1.94E-03	**4.76E-08**	**8.89E-10**
FOXA1	**0.098**	**0.1565**	**0.1872**	**2.44E-77**	**3.75E-194**	**5.07E-277**
AP1 (cFos+cJun)	**0.0218**	**0.0375**	**0.0606**	**3.35E-05**	**1.69E-07**	**1.20E-30**

Из таблицы видно, что все модификации открытого хроматина - H3K4me1, H3K9ac, FAIRE, H3K4me3, H3K14ac имеют значимую корреляцию с высотой пиков ChIP-seq связывания ERα (выделено жирным шрифтом). В то же время контрольное секвенирование, и маркеры модификаций H3K9me3 и H3K27me3 не показывают значимой корреляции (небольшое исключение показано для размеров ChIP-seq библиотек больше 5 миллионов прочтений, но и коэффициент корреляции имеет знак минус). Таким образом, численно показана значимая положительная корреляция между маркерами открытого хроматина и связыванием ERα в ChIP-seq эксперименте,

присутствующая как для полного числа прочтений (размера библиотек), так и для уменьшенного размера. При этом наибольшую корреляцию показывает маркер H4K4me1. Наблюдается также значимая положительная корреляция со связыванием других транскрипционных факторов - FOXA1 и AP1.

Поскольку при сравнении профилей ChIP-seq для сайтов связывания ERα в геноме до сих пор были показаны только общие меры зависимостей, а число прочтений в профилях модификаций хроматина могло быть очень небольшим (1-2 прочтения на сайт), встает вопрос, действительно ли эти районы обогащены числом прочтений по сравнению со средним значением в геноме для сравниваемой библиотеки ChIP-seq. Районы модификаций гистонов могут быть очень протяженными, форма пиков ChIP-seq достаточно широкая, может захватывать от одной до нескольких нуклеосом и простираться от 150 до 500 нт в геноме. Будем считать, что для модификаций хроматина рассматриваемый участок профиля ChIP-seq обогащен числом прочтений (имеет пик), если число прочтений в данном участке превышает среднее по геному минимум в два раза. Таким образом, можно долю более высоких пиков модификаций хроматина в геномном окружении сайтов связывания ERα. Следующая гистограмма показывает присутствие пиков профилей модификаций хроматина в сайтах связывания ERα, разделенных по квартилям, и связывание в контрольных районах генома - случайных локусах и в сайтах, имеющих консенсус элемента ERE, но не связанных с ERα ни в одном из ChIP экспериментов, как описано в предыдущей главе (рис. 4.14).

Рисунок 4.14 подтверждает следующие эффекты: маркеры активации экспрессии статистически обогащены на сайтах связывания, есть тренд ассоциации маркеров от первой квартили к последней. Сайты, содержащие мотивы ERE, но не связанные в ChIP-seq эксперименте, также не имеют маркеров модификаций гистонов, как и случайные геномные сайты без мотивов.

Открытость хроматина в районе локализации сайтов, определяющая их доступность к ТФ, может кодироваться не только особенностями нуклеотидного контекста, но определяться функциональным состоянием хроматина, зависящим от модификации гистонов.

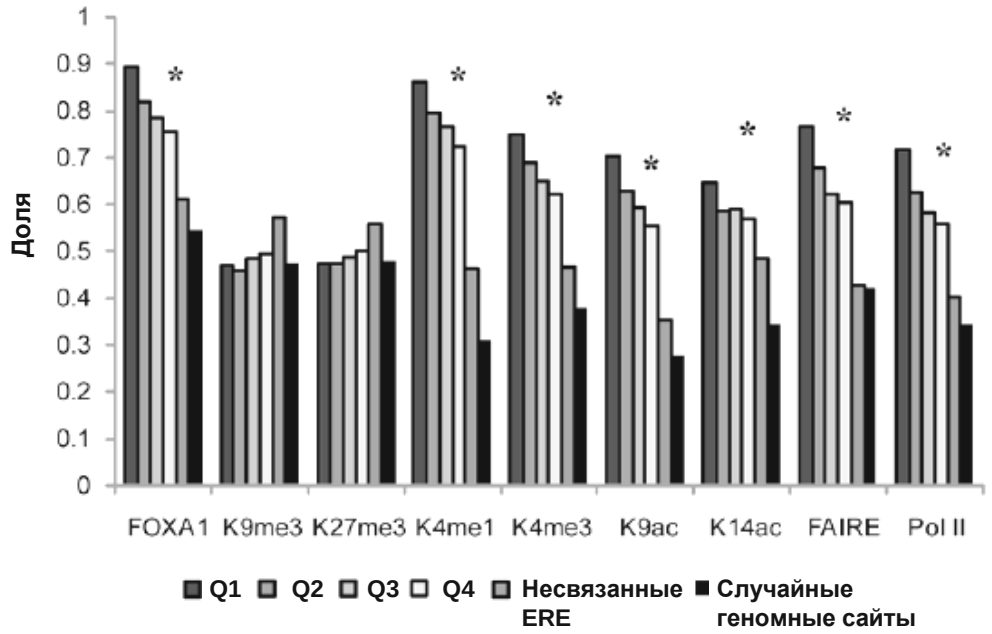


Рис. 4.14. Гистограмма ассоциации сайтов связывания ERα в геноме человека с пиками профилей ChIP-seq модификаций хроматина. Сайты разделы по группам - квартилям Q1-Q2-Q3-Q4. Ось ординат показывает долю сайтов ERα имеющих число прочтений ChIP-seq из библиотек маркеров хроматина в участке +/-250 нт от центра ERα ChIP-seq пика, превышающее как минимум в два раза число прочтений в среднем по геному (т.е. имеющих пики профиля модификаций хроматина). Звездочка (*) показывает статистическую значимость различия этой доли для четырех квартилей сайтов связывания ERα по сравнению с несвязанными геномными сайтами.

4.4 Предсказание сайтов связывания в геноме человека с помощью компьютерной модели, учитывающей состояние хроматина

Разработана компьютерная программа для предсказания сайтов ERα в геноме человека с помощью компьютерной модели, учитывающей состояние хроматина по ChIP-seq данным описанных выше модификаций гистонов, связывания РНК-полимеразы II, FAIRE, связывания FoxA1 [13]. Были подготовлены контрастные выборки - сайты связывания ERα и сайты, не связанные с ERα. В качестве последних использовались сайты, содержащие мотив ERE, но не отмеченные ни в ChIP-seq, ни в опубликованных ранее ChIP-PET и ChIP-chip экспериментах. Использовались также случайные геномные участки (около 800 тысяч участков), выбранные в хромосомных координатах с помощью датчика случайных чисел. Для построения распознающей функции и выбора оптимальной комбинации характеристик использовались стандартные программы среды R.

Рассматривались две основные задачи предсказания (дискриминации) сайтов связывания ERα в геноме человека: 1) связанные сайты против случайных дистальных геномных участков и 2) связанные сайты против несвязанных участков проксимальных

промоторов. Для решения каждой задачи последовательно использовалась одна, две, три, и так далее, характеристик (N=1, 2, 3..), и выбиралась лучшая комбинация N характеристик. Использовалась 5-кратная проверка качества распознавания при случайном разделении выборок на обучение и контроль (70% и 30% всех данных).

На рисунке 4.15 показаны кривые ошибок ROC для предсказания связывания ERα по данным секвенирования модификаций хроматина (в различных комбинациях факторов) в геноме человека. Каждая кривая представляет собой предсказание с помощью линейной комбинации указанных геномных характеристик при варьировании порога распознавания.

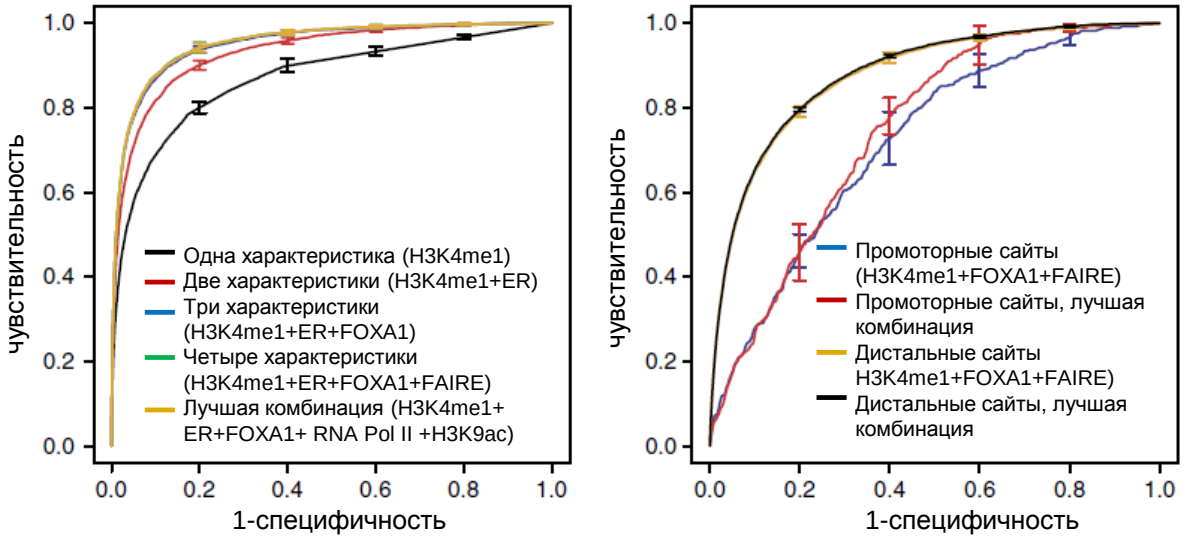


Рис. 4.15. Предсказание связывания ERα по данным секвенирования модификаций хроматина (в различных комбинациях факторов) в геноме человека. Каждая кривая ошибок представляет собой предсказание с помощью линейной комбинации геномных характеристик [13]. Левая панель - предсказания для всех непромоторных сайтов ERα в геноме. Правая панель - предсказание для промоторных и дистальных сайтов по отдельности.

Значения усреднены для пяти циклов обучения-распознавания, «усики» показывают стандартное отклонения для компьютерных циклов расчета. На левой панели рисунка 4.15 линии представляют кривые ROC для логистических регрессионных моделей (лучшие по точности модели из одной, двух, трех, четырех и шести характеристик), которые разделяют удаленные сайты связывания ERα (14 338 сайтов) для клеток MCF-7 и случайные геномные районы (820 000 участков) в геноме человека. Лучшей характеристикой для распознавания является модификация гистона H3K4me1, известная как маркер энхансерных сайтов.

Правая панель показывает качество лучших моделей для трехпараметрических моделей, использующих геномные характеристики FOXA1, H3K4me1 и FAIRE, полученные с помощью секвенирования, для разделения между ER-связанными в ChIP-

seq в клетках MCF-7 и несвязанными участками, содержащими мотив ERE (промоторные и дистальные сайты связывания показаны раздельно). Видно, что промоторные сайты связывания ERα распознаются несколько хуже, чем дистальные сайты. По-видимому, из-за насыщенности профилей секвенирования ChIP-seq в промоторных районах дальнейшее увеличение точности предсказания за счет полногеномных характеристик затруднительно.

В качестве характеристики использовалась собственно компьютерное предсказание мотива связывания ERα. Мотив связывания, использованный в распознавании, был рассчитан с помощью оптимизированной весовой матрицы. Матрица оптимизировалась по методу TherMoS (Thermodynamic Modeling of chip-Seq) [13].

Предсказание связывания ERα по данным секвенирования модификаций хроматина (в различных комбинациях факторов) в геноме человека при обучении модели на данных из линии MCF-7 и тестировании предсказания на данных клеточной линии T47D также показывает высокий результат площади под кривой (см. рисунок 4.16) [13].

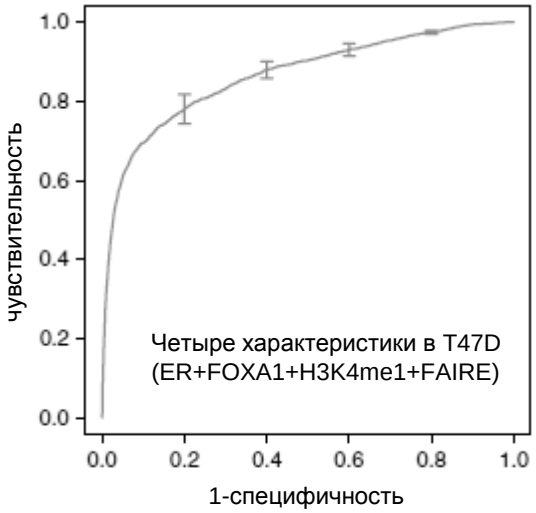


Рис. 4.16. Предсказание связывания ERα в геноме человека по данным секвенирования модификаций хроматина в клеточной линии MCF-7 data (в модели логистической регрессии комбинации факторов: связывания ERα в MCF-7, связывания FOXA1, модификации H3K4me1 и FAIRE) при проверке связывания в клеточной линии T47D [13].

Использовались не все геномные характеристики, а только те, для которых ChIP-seq эксперимент был выполнен в обеих клеточных линиях (см. Таблицу 4.1). Таким образом, возможно предсказание сайтов связывания ERα, не зависящее от клеточной линии.

Таблица 4.3 представляет лучшую комбинацию N характеристик для задачи 1 дискриминации сайтов (связанные сайты против случайных дистальных геномных участков) используя 5-кратную проверку.

Таблица 4.3

Лучшая комбинация N характеристик для задачи 1 дискриминации сайтов (связанные сайты против случайных дистальных геномных участков) используя 5-кратную проверку

N	**Наиболее частая комбинация лучших характеристик для распознавания**	**Среднее значение ROC-AUC**
1	* H3K4me1	0.870
2	* H3K4me1+ER мотив	0.929
3	* H3K4me1+ER мотив+FOXA1	0.949
4	* H3K4me1+ER мотив+FOXA1+FAIRE	0.952
5	H3K4me1+ER мотив+FOXA1+FAIRE+RNA Pol II	0.953
6	H3K4me1+ER мотив+FOXA1+FAIRE+RNA Pol II+H3K9ac	0.953
7	H3K4me1+ER мотив+FOXA1+FAIRE+RNA Pol II+H3K9ac+ H3K4me3	0.953
8	H3K4me1+ER мотив+FOXA1+FAIRE+RNA Pol II+H3K9ac+ H3K4me3+H3K27me3	0.950
9	H3K4me1+ER мотив+FOXA1+FAIRE+RNA Pol II+H3K9ac+ H3K4me3+H3K9me3+ H3K14ac	0.948
10	H3K4me1+ER мотив+FOXA1+FAIRE+RNA Pol II+H3K9ac+ H3K4me3+H3K9me3+ H3K14ac+H3K27me3	0.946
11	H3K4me1+ER мотив+FOXA1+FAIRE+RNA Pol II+H3K9ac+ H3K4me3+H3K9me3+ H3K14ac+H3K27me3+cJun	0.945

Из таблицы видно, что при увеличении числа характеристик распознающей функции более 7 точность перестает расти (максимальное значение 0.953). 4 или 5 характеристик уже достаточно для высокой точности распознавания, отличающейся на сотую долю единицы. Отметим, что маркер модификации гистонов H3K4me1 всегда входит в число лучших характеристик распознавания сайтов во всех комбинациях.

В следующей таблице, представляющей наборы характеристик для предсказания сайтов (дискриминации по группам) в задаче 2 (связанные сайты против несвязанных участков проксимальных промоторов) использовались оценки ROC-AUC аналогично предыдущей таблице.

Также маркер модификации гистонов H3K4me1 является лучшей распознающей характеристикой, хотя общая точность несколько ниже, чем для предсказания дистальных (не промоторных) сайтов связывания в геноме.

Таблица 4.4

Лучшая комбинация N характеристик для задачи 2 дискриминации сайтов (связанные сайты против несвязанных участков проксимальных промоторов).

N	**Наиболее частая комбинация лучших характеристик для распознавания**	**Среднее значение ROC-AUC**
1	ER мотив	0.833
2	ER мотив+H3K4me1	0.900
3	ER мотив+H3K4me1+FOXA1	0.912
4	ER мотив+H3K4me1+FOXA1+RNA Pol II	0.916
5	ER мотив+H3K4me1+FOXA1+cFos+FAIRE	0.918
6	* комбинация 6 характеристик из 12	0.920
7	* комбинация 7 характеристик из 12	0.921
8	* комбинация 8 характеристик из 12	0.922
9	* комбинация 9 характеристик из 12	0.922
10	* комбинация 10 характеристик из 12	0.922
11	* комбинация 11 характеристик из 12	0.923

Примечание: * Не было единственной комбинации этих N характеристик из 12 возможных, устойчиво показывающей лучшую точность распознавания.

Следующий рисунок показывает графики распознавания сайтов связывания ERα после активации клеток MCF-7 эстрадиолом (E2) - оценку площади под кривой - для лучших отдельных характеристик, настроенных на ChIP-seq характеристики клеток в растворителе (DMSO), без активации экспрессии ERα.

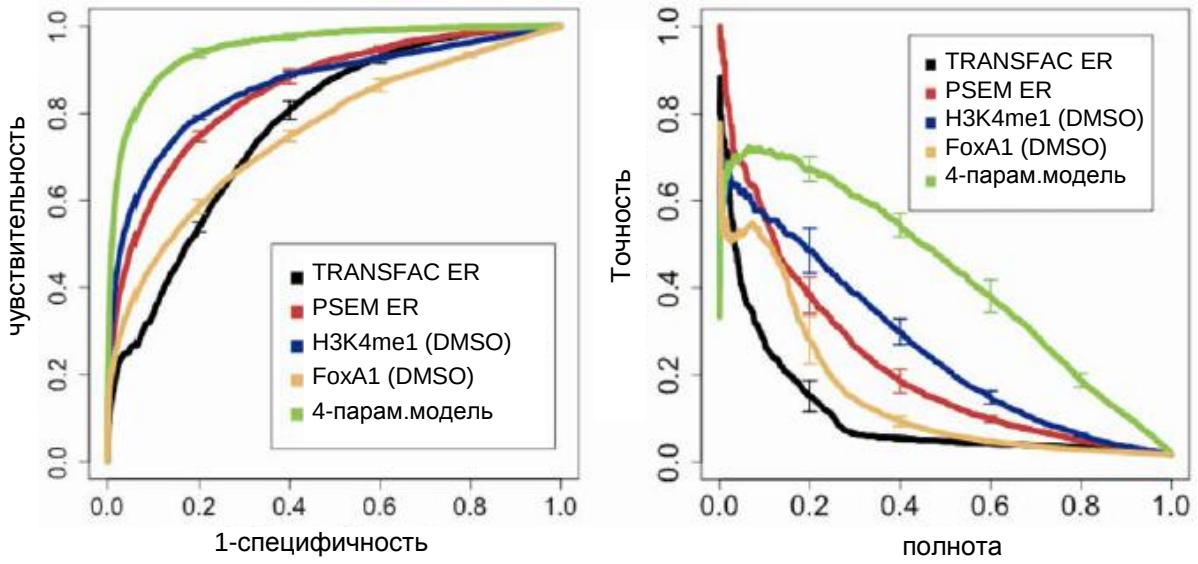


Рис. 4.17. Сравнение точности предсказания связывания ERα в геноме с помощью различных комбинаций факторов. Левая панель представляет кривые ошибок предсказания ROC, правая панель - кривые точности и полноты («Precision-recall»).

Библиотеки ChIP-seq для клеток MCF-7 были нормализованы (сокращены) до 7 миллионов прочтений каждая. Были построены кривые точности предсказания ROC и «precision-recall» (точность-полнота) для предсказания с помощью позиционных

весовых матриц, построенным по базе данных TRANSFAC, позиционных энергетических матриц, построенных по данным связывания ERα (TherMoS PSEM), профилям связывания ChIP-seq модификаций гистонов H3K4me1 (высотам пиков профиля), и 4-параметрической логистической регрессионной модели использующей совместно матрицу предсказания PSEM и профили ChIP-seq для H3K4me1, FOXA1, FAIRE. Значения были усреднены для 5 циклов обучения и контроля (размах значений показан на кривых).

Из рисунка 4.17 видно, что наибольшую точность из индивидуальных характеристик дает H3K4me1. Весовые матрицы для распознавания мотива связывания ERα, как стандартная матрица базы данных TRANSFAC, так и оптимизированная для данных ChIP-seq позиционно-специфическая весовая матрица PSEM работают не так хорошо.

Отметим, что при использовании данных ChIP-seq, нормализованных до 7 миллионов прочтений для каждой из используемых в предсказании библиотек, показан такой же высокий результат точности по площади под кривой ROC-AUC (0.952 для модели, использующей те же четыре параметра - матрицу предсказания PSEM и профили ChIP-seq для H3K4me1, FOXA1, FAIRE) как и результат, полученный для использования полных данных секвенирования без нормализации [13].

На следующем рисунке 4.18 представлена кривая ошибок ROC предсказания связывания ERα для сайтов, находящихся в промоторных районах (интервал [-500..-1] нуклеотидов относительно старта транскрипции гена RefSeq) в клетках MCF-7.

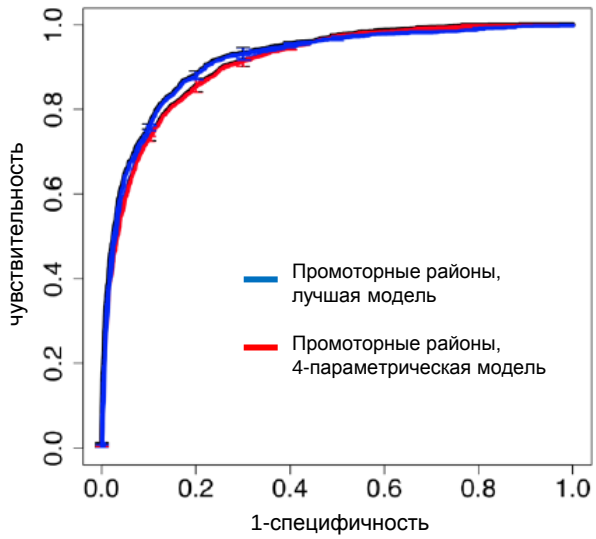


Рис. 4.18. Сравнение точности предсказания связывания ERα для промоторных районов с помощью оценок площади под кривой для 4-параметрической модели и модели, использующей все характеристики.

Модель, использующая 4 параметра, включающие скор аффинности TherMoS, сигнал высоты ChIP-seq пиков для H3K4me1, FOXA1 и FAIRE достигает значения

ROC-AUC 0.915 (подтверждено 5-кратным выборочным тестированием). Лучшая модель, включающая все характеристики, достигает величины ROC-AUC 0.9225. Рисунок показывает незначительное увеличение точности при использовании лучшей модели по сравнению с 4-параметрической моделью.

Рисунок показывает распределение высоты пика ChIP-seq маркера H3K4me1, FAIRE и связывания FOXA1 на ССТФ ERα для двух позитивных (MCF-7, T47D) клеточных линий рака молочной железы человека.

На следующем рисунке 4.19 показан средний размер пика ChIP-seq маркера модификации гистона H3K4me1, секвенирования по методу FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements), профиля связывания транскрипционного фактора FOXA1 в культуре клеток MCF-7 (верхний ряд панелей, A–C) и культуре клеток T47D (нижний ряд, панели D–F) соответственно, для сайтов связывания ERα.

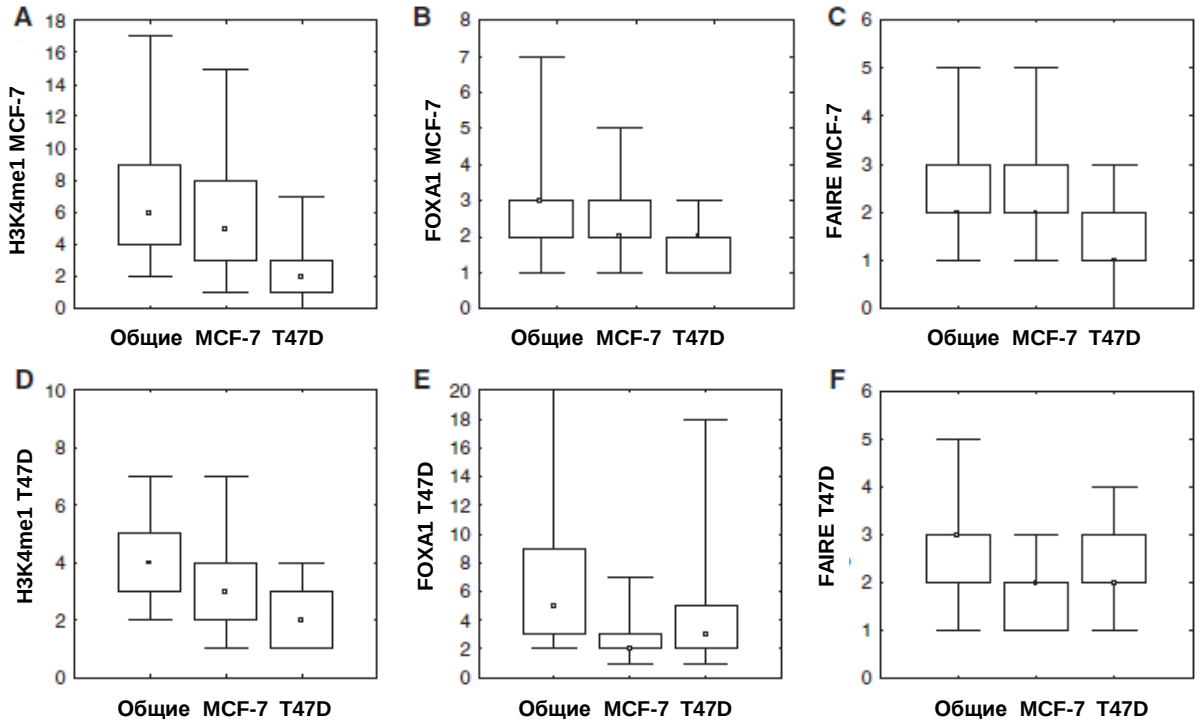


Рис. 4.19. Распределение высоты пика ChIP-seq маркеров на сайтах связывания ERα для двух клеточных линий рака молочной железы. Показан средний размер пика ChIP-seq профиля для маркера модификации гистона H3K4me1, секвенирования по методу Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE), профиля связывания транскрипционного фактора FOXA1 в культуре клеток MCF-7 (панели A–C) и культуре клеток T47D (панели D–F) соответственно, для сайтов связывания ERα.

Расчет числа прочтений ДНК в этих экспериментах ChIP-seq выполнялся для района ±250 нт вокруг центра связывания сайта ERα в трех вариантах: для сайтов, общих для клеточных линий MCF-7 и T47D, и специфичных только к одной из линий. Показаны данные ChIP-seq перед активацией клеток эстрадиолом (E2) с нормализацией числа прочтений к размеру ChIP библиотек. Сайты связывания ERα общие для двух

клеточных линий («общие») значимо обогащены маркерами модификаций хроматина как в клетках MCF-7, так и в клетках T47D.

В целом видна ассоциация сайтов, специфичных для клеточной линии, с маркерами хроматина, специфичными для той же линии. Общие сайты и сайты, специфичные для MCF-7, обогащены маркерами (A) H3K4me1 (B) FOXA1 и (C) FAIRE в клеточной линии MCF-7, в то время как сайты связывания ERα, специфичные для линии T47D, имеют значительно меньшую высоту пика ChIP-seq, чем сайты, специфичные для MCF-7. Панели (D–F) показывают обогащение ChIP-seq фрагментов для маркеров H3K4me1, FOXA1 и FAIRE соответственно, в клетках линии T47D, где, как видно из распределений, маркеры FAIRE и FOXA1 в специфичных для T47D сайтах связывания ERα имеют значительно более высокий сигнал в сравнении с сайтами, специфичными для линии MCF-7.

Таблица 4.5 показывает оценки статистической значимости различий в уровне ChIP-seq сигнала маркеров модификаций гистонов для сайтов ERα, специфичных к клеточным линиям MCF-7 и T47D. Представлены сайты, специфичные для клеток линии MCF-7 и для клеток линии T47D, а также сайты, общие для обеих линий. Показана Z статистика и соответствующие уровни значимости (P-value) для критерия Манна-Уитни (Mann-Whitney U test).

Таблица 4.5

Статистическая значимость различий в уровне маркеров модификаций гистонов для сайтов ERα, специфичных к клеточным линиям MCF-7 и T47D

	Z статистика (уровень значимости *P*) разницы в маркерах хроматина для групп сайтов связывания ERα		
	Общие сайты против MCF-7	**MCF-7 против T47D**	**Общие сайты против T47D**
Сигнал аффинности к ERα	25.88(<1E-16)	17.61(<1E-16)	28.71(<1E-16)
ChIP-seq в клетках MCF-7			
FOXA1	0.54 (>0.1)	18.07 (<1E-16)	15.35 (<1E16)
H3K4me1	26.46 (>0.1)	25.74007 (<1E16)	23.08011(<1E16)
FAIRE	0.148(>0.1)	24.06 (<1E16)	20.75 (<1E16)
ChIP-seq в клетках T47D			
FOXA1	0.347 (>0.1)	-7.53 (4.95E-14)	-6.77 (1.29E-11)
H3K4me1	0.739 (>0.1)	26.45 (<1E-16)	22.23 (<1E-16)
FAIRE	0.238 (>0.1)	-12.34 (<1E-16)	-10.56 (4.42E-26)

Из таблицы видна высокая статистическая значимость различий маркеров хроматина между клеточными линиями и в сравнении с общими, более сильными и консервативными сайтами связывания ERα.

4.5. Общая зависимость доступности ССТФ от состояния хроматина опосредована присутствием нуклеосом на ДНК

Присутствие нуклеосом на участках ССТФ может быть оценено с помощью компьютерных моделей или экспериментальных данных секвенирования. Для общей оценки доступности участков, содержащих сайты связывания, в масштабе генома с учетом локальной открытости хроматина рассмотрим предсказательную модель предрасположенности нахождения нуклеосом для сайтов связывания белка REST, определенных в геноме мыши [280].

Доступная ранее в базах данных информация [255, 268] о последовательностях ДНК, для которых экспериментально подтверждено существование нуклеосомной упаковки (около 200 последовательностей), мала по сравнению с имеющимися полногеномными данными. Тем не менее, использование таких обучающих данных и марковской модели порождения последовательностей, связывающихся с нуклеосомой, позволяет рассмотреть проблему нуклеосомного кода с новой точки зрения, что было представлено в работе автора [50]. Действительно, слабый контекстный код позиционирования нуклеосом предполагает вырожденность (совершенно различные последовательности ДНК способны к взаимодействию гистоновым октамером и образованию нуклеосом), слабость контекстных сигналов, и отсутствие их чёткой локализации. В марковских моделях порождения символьных последовательностей не используются конкретные сигналы, и в то же время оценивается зависимость появления символа от локального контекста. Таким образом, как статистическая модель, марковская модель соответствует теоретическим предположениям о нуклеосомном коде. Использовалась марковская модель с переменной памятью VMM (Variable Memory Markov model) и соответствующая компьютерная программа, рассчитывающая для обучающей выборке оптимальный набор контекстов (длины от 2-х до 8-ми нуклеотидов), по зависимостям от которых можно описать нуклеотидные последовательности в выборке.

На рисунке показан рассчитанный с помощью программы VMM [50] усредненный профиль потенциала нуклеосомной упаковки ДНК для района локализации 685 сайтов, взаимодействующих с репрессором транскрипции REST (RE1-silencing transcription factor), экспериментально выявленных в геноме мыши [280] с помощью технологии ChIP-chip. В этом случае потенциал нуклеосомной упаковки минимален в области локализации сайта. Это означает, что сайты данной группы локализуются в участках с ослабленной нуклеосомной упаковкой, т.е. белок REST имеет облегченный доступ к ДНК. Усредненный профиль нуклеосомной упаковки для

147 сайтов, имеющих более низкое сродство к REST, также имеет минимум в месте локализации сайтов, однако, он менее выражен, чем для сайтов первой группы.

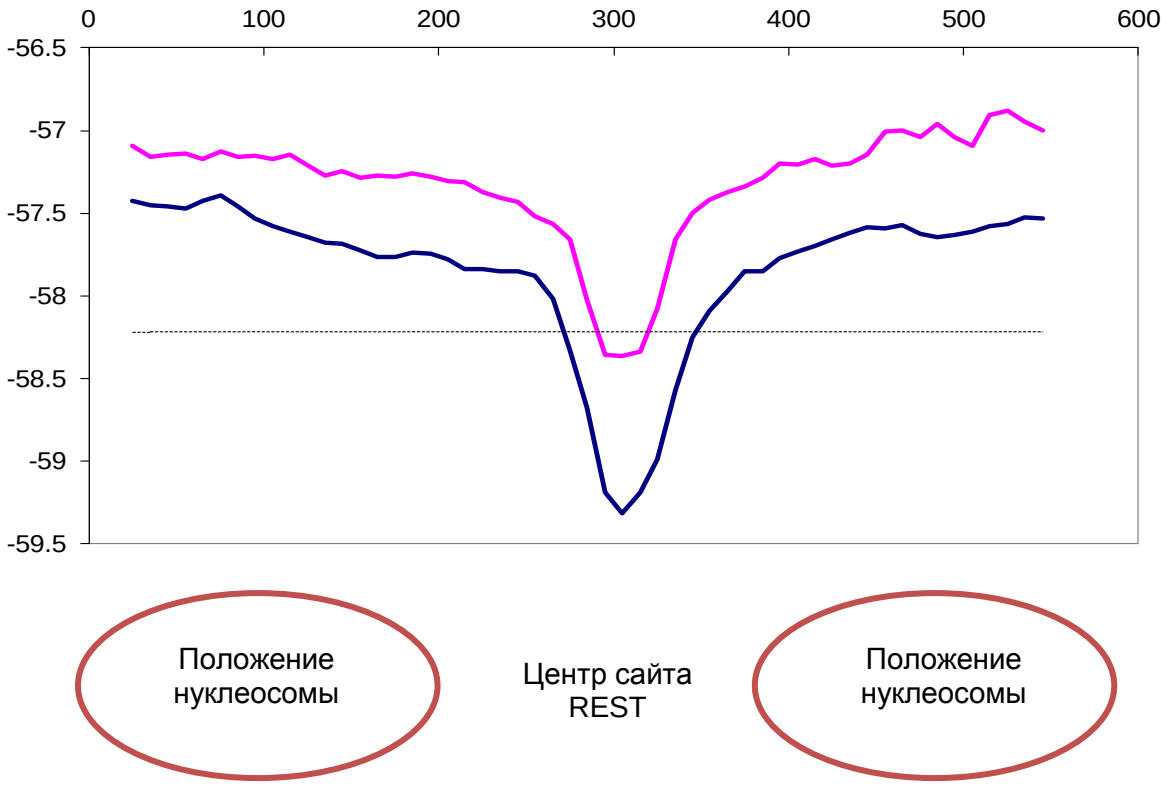


Рис. 4.20. Усредненный профиль потенциала нуклеосомной упаковки ДНК, рассчитанный с помощью программы VMM для фазированной выборки нуклеотидных последовательностей, содержащей, сайты связывания транскрипционного фактора REST (RE1-silencing transcription factor).

Этот пример показывает, что контекст, содержащий сайты связывания REST, содержит перекрывающиеся генетические сообщения, записанные в двух разных кодах, один из которых имеет отношение к нуклеосомной упаковке ДНК, а второй – к определению локализации REST и величины его сродства к ДНК.

Для корректного сравнения экспериментальных данных по состоянию хроматина и присутствия ССТФ необходимо параллельное проведение экспериментов ChIP-seq на том же клеточном материале, как для определения локализация сайтов связывания, так и для определения сайтов или профиля представленности нуклеосом. Состояние хроматина может быть определено не только прямым секвенированием нуклеосом, но и секвенированием ДНК, связанным с определенными модификациями нуклеосом, точнее входящих в их состав гистонов. Именно такие характеристики и являются определяющими для предсказания ССТФ, как было показано в предыдущем разделе.

4.6. Заключение к Главе. Общая проблема предсказания сайтов связывания на основе данных о модификациях хроматина

В данной Главе аргументировано следующее положение, выносимое на защиту:

Расположение сайтов связывания транскрипционного фактора ERα в геноме человека положительно ассоциировано с районами метилирования и ацетилирования гистонов нуклеосом H3K4me3, H3K4me1, H3K9ac и H3K14ac. Разработан компьютерный алгоритм для предсказания сайтов связывания ERα в геноме по ChIP-seq маркерам состояния хроматина; показана высокая точность предсказания с помощью этой модели.

Проблема предсказания сайтов связывания в полногеномном эксперименте может быть обобщена для других транскрипционных факторов и полногеномных технологий, основанных на иммунопреципитации хроматина и секвенировании [13, 44]. Обычно такие данные анализируются по отдельности из-за непроработанности общей теоретической модели полногеномного распределения регуляторных районов и необходимости комбинирования данных различных технологий высокопроизводительного секвенирования. Тем не менее, многие проблемы анализа данных такого рода изучались ранее в классической теории обработки сигналов - можно оптимизировать отношение сигнала к шуму (сигнала связывания профиля ChIP-seq), объединить геномные характеристики в оптимальной линейной комбинации, чтобы улучшить точность предсказания (максимизировать площадь под кривой ROC-AUC). Такое объединение характеристик может включать различные типы данных секвенирования, как гистонов, так и транскрипционных факторов по методам ChIP-seq, FAIRE-seq и другим сигналам секвенирования подобного рода. Такой подход позволяет увеличить точность предсказания сайтов связывания, поскольку нуклеотидные мотивы и паттерн связывания в промоторе коррелируют с модификациями гистонов.

В целом результаты, приведенные в данной главе, позволяют обосновать следующие выводы:

Расположение сайтов связывания транскрипционного фактора ERα в геноме человека положительно ассоциировано с районами метилирования и ацетилирования гистонов нуклеосом H3K4me3, H3K4me1, H3K9ac и H3K14ac. Разработан компьютерный алгоритм для предсказания сайтов связывания ERα в геноме по ChIP-seq маркерам состояния хроматина; показана высокая точность предсказания с помощью этой модели.

Глава 5. ХРОМОСОМНЫЕ КОНТАКТЫ И РЕГУЛЯЦИЯ ТРАНСКРИПЦИИ В ГЕНОМЕ ЧЕЛОВЕКА

5.1. Введение к Главе 5. Проблема исследования хромосомных контактов

Глава 5 содержит материалы исследования хромосомных контактов, полученные с помощью массового параллельного секвенирования по методу ChIA-PET, и представленные в работах автора [12, 21].

Решаемая задача:

Изучение роли хромосомных контактов в регуляции транскрипции генов человека на моделях РНК-полимеразы II и транскрипционного фактора ERα на основе компьютерного анализа полногеномных данных ChIP-seq и ChIA-PET.

В разделах данной главе описаны принципы построения карт хромосомных контактов по данным ChIA-PET с помощью компьютерных программ, показаны примеры применения для контактов, опосредованных связыванием транскрипционного фактора ERα, показаны модели контактов для РНК-полимеразы II.

Далее, в настоящей Главе показана ассоциация участков хромосомных контактов с регуляторными районами транскрипции, сайтами связывания транскрипционных факторов и участками открытого хроматина, размеченными по технологии ChIP-seq в геноме человека [12]. Исследована информация о хромосомных контактах, опосредованных комплексом РНК-полимеразы II, и рецептором эстрогена ERα, полученных с помощью метода ChIA-PET, показаны приложения по интеграции геномных карт контактов с сайтами связывания транскрипционных факторов и районами модификаций гистонов, определенных с помощью. Использование методов определения контактов через секвенирования сближенных фрагментов ДНК позволило выявить ряд важных особенностей пространственной организации генома соматических клеток [64].

5.1. Принципы построения карт хромосомных взаимодействий и компьютерные модели

Карты контактов на хромосомах в экспериментах по определению структуры хроматина с помощью секвенирования строят на основе таблиц парных контактов секвенированных фрагментов. Для метода ChIA-PET было введено понятие петли хромосомы, когда участок между контактами формирует петлю или несколько петель.

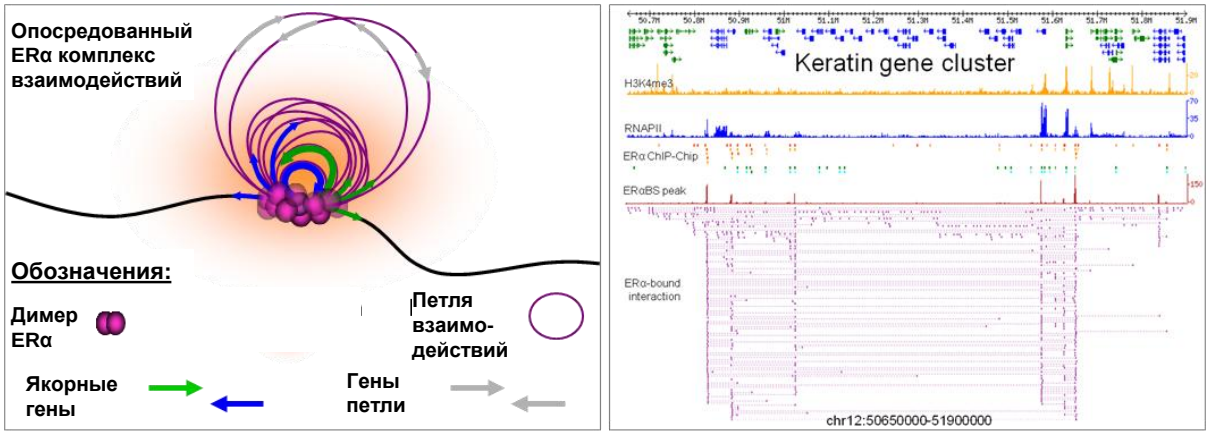


Рис. 5.1. Схема представления контактов в хромосоме – двумерная карта, «тепловая карта» плотности контактов.

Данные хромосомных контактов подтверждались с помощью экспериментов по технологии 3C (Chromosome conformation capture) [484] и флюоресцентной гибридизации in situ (FISH). При помощи метода Hi-C ранее независимо было подтверждено наличие в ядре хромосомных территорий [482]. Отмечается, что межхромосомные контакты очень динамичны, то есть в различных клетках одной и той же клеточной популяции распределение контактов может значительно отличатся [22]. Показано, что районы активно транскрибируемых генов чаще участвуют в межхромосомных контактах. По-видимому, этот факт объясняется выделением петель хромосом в область концентрации белков транскрипционного аппарата (так называемые «фабрики транскрипции»), где становится возможным контакт с другим активным районом [64].

В последние годы с использованием методов Hi-C, ChIA-PET и TCC получены новые знания об особенностях трехмерной архитектуры (укладки) генома человека в интерфазном ядре [483]. Методы основаны на лигировании сближенных в пространстве фрагментов хромосомной ДНК с последующим высокопроизводительным параллельным секвенированием и картированием (компьютерным выравниванием) секвенированных фрагментов на полный геном. На основе этих данных для генома человека сконструированы карты сближенности фрагментов ДНК с разрешением до сотен нуклеотидов и построены модели пространственного расположения хромосом в ядре [12, 482], указывающие на наличие хромосомных территорий, образованных пространственно сближенными участками ДНК. Хромосомные территории – сближенные в пространстве ядра районы хромосом, которые обеспечивают формирование транскрипционных фабрик и транскрипционных доменов, содержащих группы коэкспрессирующихся генов, называемые "хромопероны", или хромосомные опероны [12], обслуживаемые общими пулами РНК-полимераз и транскрипционных

факторов. Здесь возникает задача изучения качественно новых кодов регуляции транскрипции, реализующихся на надхроматиновом уровне, таких как транскрипционные домены [483].

Использование методов определения контактов через секвенирование сближенных фрагментов ДНК позволило выявить ряд важных особенностей пространственной организации генома соматических клеток. Было показано, что укладка длинных молекул ДНК в очень небольшом объеме ядра достигается за счет упаковки по модели «фрактальной глобулы». При таком способе упаковки близко расположенные участки молекулы ДНК предпочтительно взаимодействуют друг с другом, формируя глобулу первого порядка. В свою очередь, близко расположенные глобулы первого порядка так же с большей вероятностью взаимодействуют друг с другом, формируя глобулу второго порядка, и т.д. Модель «фрактальной глобулы» позволяет объяснить компактизацию длинной молекулы ДНК и образования при этом большого количества нераспутываемых узлов, которые неизбежно возникают при случайной, стохастической компактизации длинного полимера. Кроме того, фрактальная модель организации хроматина объясняет, как быстро компактизовывать и декомпактизовывать отдельные участки молекулы ДНК, тем самым обеспечивая выполнение ее функции [482].

Проблема анализа трехмерной структуры генома активно исследуется различными методами, использующими геномное секвенирование. Метод Hi-C позволяет реконструировать карту пространственных взаимодействий ДНК в ядре клетки, однако разрешение подобной карты сильно зависит от глубины секвенирования ДНК библиотеки. Метод ChIA-PET позволяет определять контактирующие участки хромосом, контакты которых опосредованы белками или белковыми комплексами. В данной работе описан компьютерный анализ экспериментов по определению контактов в геноме человека, опосредованных белком - рецептором эстрогена [21] и комплексом РНК-полимеразы II [12].

5.2. Анализ трехмерной структуры генома через секвенирование. ChIA-PET, Hi-C технологии

Был разработан метод анализа взаимодействий хроматина с помощью секвенирования парных концов ДНК (Chromatin Interaction Analysis by Paired-End-Tag sequencing - ChIA-PET) для полногеномного исследования взаимодействий хроматина, связанного со специфическими белковыми факторами [21]. С помощью хроматин-

иммунопреципитации исследуемый белковый фактор вместе с ассоциированными с ним фрагментами ДНК, извлекается из клеточных ядер. Выполняется последующее лигирование в растворе фрагментов ДНК, закрепленных вместе в отдельном комплексе хроматина, включая фрагменты, которые могли находится удаленно по хромосоме и даже на разных хромосомах. Далее строится карта контактов и исследуется ассоциация регуляторных районов через их нелинейные взаимодействия.

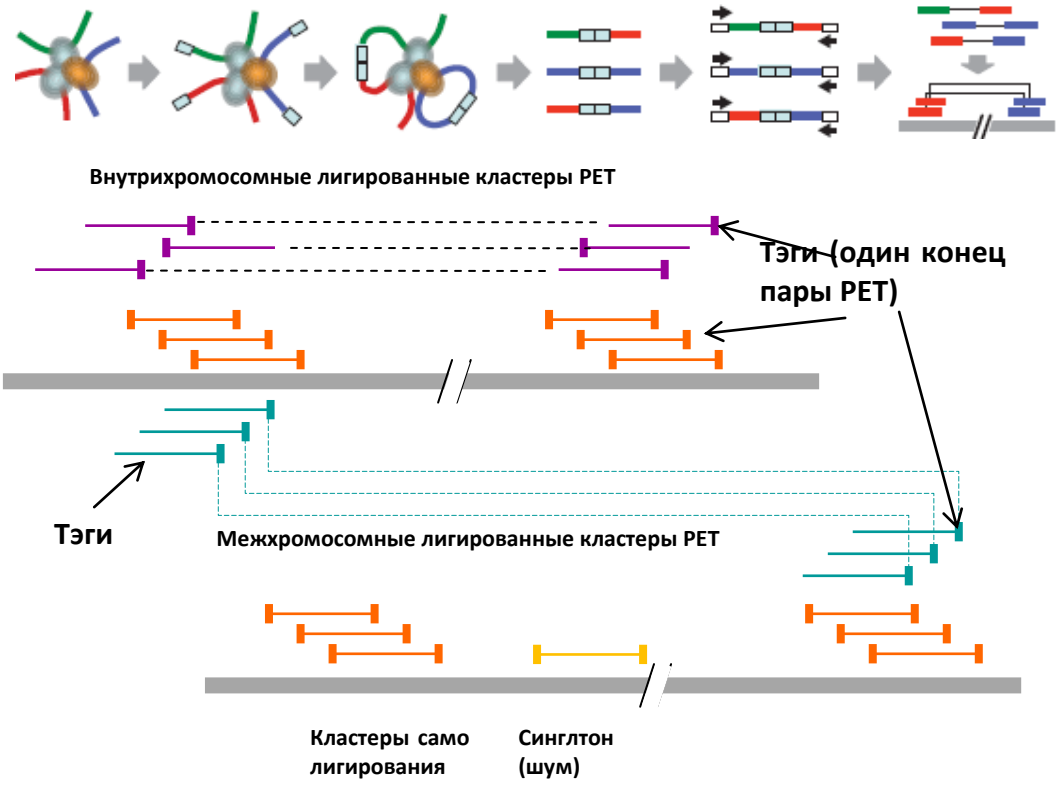


Рис. 5.2. Диаграмма метода ChIA-PET. Фрагменты ДНК из разделенных ультразвуком, прошедших иммунопреципитацию хроматиновых комплексов были обработаны через лигирование линкеров (на свободные концы ДНК), лигирование сближенных фрагментов, получены парные концы (PET). Далее ДНК секвенируется и картируется на референсную последовательность генома для выявления взаимодействующих в пространстве участков хромосом [21].

Исследована информация о хромосомных контактах, опосредованных комплексом РНК-полимеразы II и рецептором эстрогена ERα, полученных с помощью нового метода ChIA-PET. Продемонстрировано, что удаленные контакты хроматина встречаются между дистальными участками связывания рецептора эстрогенов ERα и промоторами генов-мишеней этого фактора.

Анализ хромосомных контактов, опосредованных комплексами РНК-полимеразы II, в интерфазных ядрах клеток человека выявил обогащенность

контактирующих участков сайтами связывания различных транскрипционных факторов. Отдельные хромосомные контакты, определенные через секвенирование, были протестированы экспериментально с помощью стандартного эксперимента 3C и FISH.

На рисунке 5.3 представлена схема расположения нуклеотидных последовательностей парных концов ChIA-PET вместе с линкерами. Рисунок показывает палиндромную последовательность линкера.

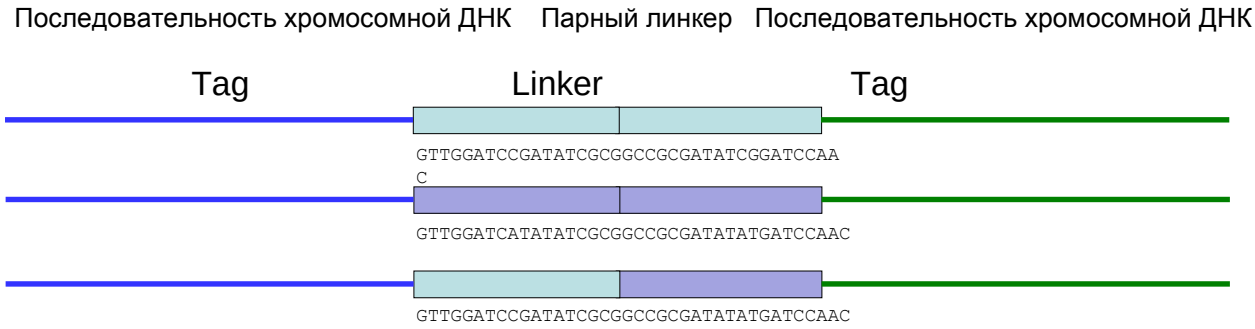


Рис. 5.3. Схема нуклеотидных последовательностей парных концов ChIA-PET.

5.3 Хромосомные контакты, опосредованные связыванием транскрипционного фактора ERα в геноме человека

Был сгенерирован большой набор данных ChIA-PET для связывания ER (идентификатор библиотеки секвенирования IHM001F) объемом около 32 миллионов парных последовательностей PET. Использовалось секвенирование парных фрагментов на оборудовании Illumina GAII (см. Таблицу 5.1) для детального анализа связывания ER и взаимодействий хроматина в обработанных эстрогеном клетках линии MCF-7 [21]. Из 4.63 миллионов уникально картированных последовательностей PET только 1.23 миллиона (27%) были кластерами самолигирования парных концов (из одного геномного района). Среди этих PET самолигирования, 16.7% сформировали перекрывающиеся кластеры PET, образуя 14 468 пиков - предполагаемых сайтов связывания ER (ошибка перепредсказания (FDR) была в пределах 0.01, число PET формирующих пик было по меньшей мере 5). Из общего числа лигированных фрагментов ДНК только 5.1% от общего числа картированных PETs были внутрихромосомными. После статистической обработки данных, исключения синглтонов PET лигирования как слабых взаимодействий или шумового сигнала секвенирования, определения кластеров PET в геноме из перекрывающихся групп парных концов PET был определен большой набор 1 451 внутрихромосомных и небольшой набор 15 межхромосомных перекрывающихся кластеров, состоящих из трех

и более PET на кластер (уровень ошибки перепредсказания FDR был оценен в пределах 0.05). При определении кластеров учитывалась коррекция на структурные амплификации в геноме линии MCF-7 (дуплицированные районы учитывались отдельно, как описано в предыдущей главе).

Таблица 5.1 представляет число секвенированных парных концов PET и кластеров различных типов.

Таблица 5.1

Число секвенированных PET и кластеров внутри- и межхромосомных контактов

Код библиотеки	Общее число PET	Число уникальных PET	Самолигирование		Внутрихромосомные контакты		Межхромосомные контакты	
			PET	кластеры PET	PET	кластеры PET	PET	кластеры PET
IHM062 (IgG)	436,248	217,708	40,847	0	11,254	0	165,607	0
IHM001F	31,828,194	4,638,633	1,249,081	14,560	234,400	1,451	3,155,152	15
IHH015F	19,590,581	6,125,099	1,841,684	6,665	348,057	3,543	3,935,358	4

Из таблицы видно, что значительная часть контактов связана с самолигированием, внутрихромосомных контактов значительно меньше. В то же время, межхромосомные контакты не подтверждаются образованием кластеров (незначительное число кластеров на весь геном). Контрольная библиотека секвенирования IHM062 с использованием неспецифического антитела иммуноглобулина IgG не дает кластеров PET. Таким образом, образование кластеров значимо, и не может быть объяснено ошибками эксперимента.

На рисунке 5.4 приведен пример сайтов связывания ERα, вовлеченных в комплексные взаимодействия (два и более контактов) для гена *FOXA1*.

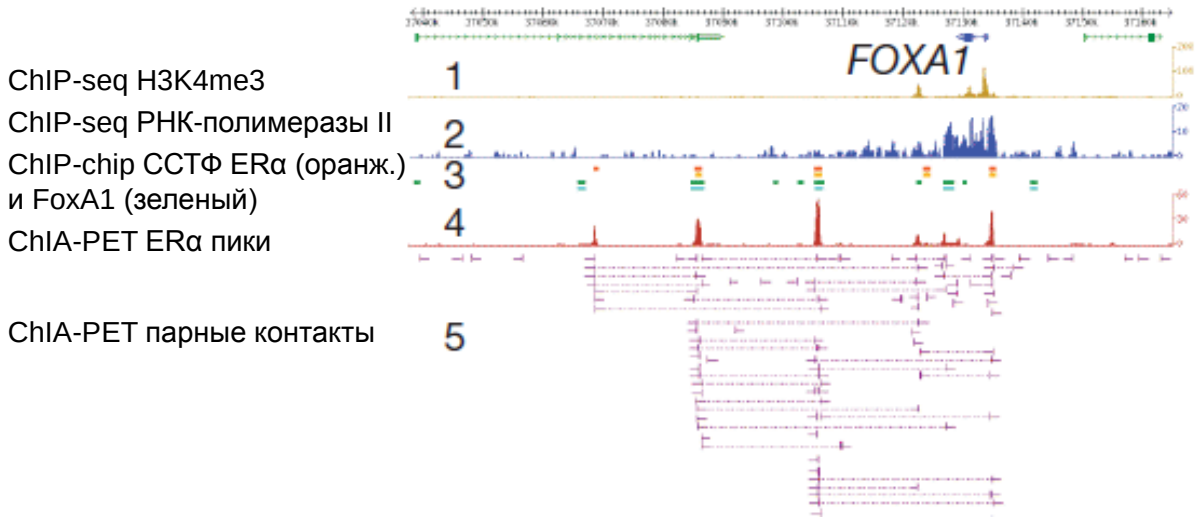


Рис. 5.4. Схема нуклеотидных последовательностей парных концов ChIA-PET [21].

Горизонтальные линии представляют геномную аннотацию: 1 - профиль ChIP-seq метилирования модификации гистонов H3K4me3, соответствующей активному хроматину; 2 - профиль ChIP-Seq для РНК-полимеразы II; 3 - районы связывания транскрипционных факторов ERα (оранжевый цвет) и FoxA1 (зеленый цвет), определенные с помощью ChIP-chip; 4 - расположение пиков ChIA-PET связывания ERα; 5 - участки контактов PET (лигирование между сайтами).

Из рисунка видно, что существует несколько контактирующих сайтов вокруг гена *FOXA1* (тонкие фиолетовые линии). Точки контактов ChIA-PET соответствуют пикам ChIP-seq связывания и сайтам связывания ER, аннотированным в геноме ранее с помощью ChIP-chip. Сам ген *FOXA1* отмечен (короткий район вверху) профилем ChIP-seq для РНК-полимеразы II, что показывает транскрипцию гена (район шире размеров сайта). Структура контактов ChIA-PET вокруг гена имеет комплексный характер и включает несколько сайтов связывания ER.

Следующий рисунок представляет геномную аннотацию и расположения так называемого дуплексного кластера ChIA-PET (два контактирующих участка) в окрестностях гена *YIF1B*.

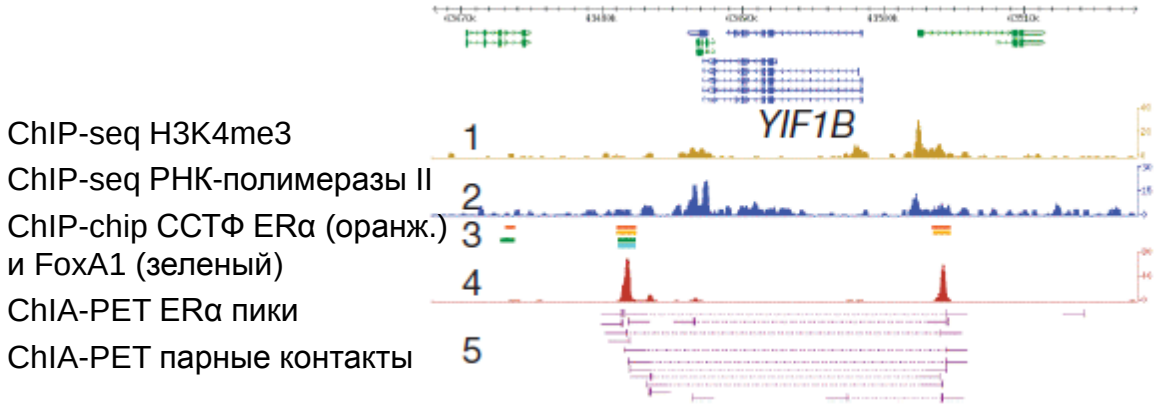


Рис. 5.5. Расположение дуплексного кластера ChIA-PET в окрестностях гена *YIF1B*.

Горизонтальные линии и цифры представляют геномную аннотацию ChIP экспериментов в локусе гена *YIF1B*, в тех же обозначениях, как и на предыдущем рисунке: 1 - ChIP-seq H3K4me, 2 - ChIP-seq для РНК-полимеразы II, и т.д.

Видно присутствие двух четко выраженных пиков связывания с контактами между ними, окружающих ген *YIF1B*.

На рисунке 5.6 представлены группы сайтов связывания ER, определенные в эксперименте ChIP-chip. Исследовалась воспроизводимость эксперимента иммунопреципитации, оцененная по перекрыванию с другим набором данных ChIA-PET (библиотека IHH015F), данными ChIP-seq и данными ChIP-chip экспериментов.

Левая панель показывает воспроизводимость как график доли числа сайтов с низким сигналом ChIA-PET (5–49 PET на сайт), перекрывающихся с перечисленными категориями ChIP-определенных сайтов из других экспериментов в зависимости от расстояния между ними. Профиль построен для расстояний от центра ChIA-PET сайта до границы других ChIP сайтов с шагом 50 нт. Использовалось 11955 сайтов данной группы ChIA-PET сайтов. Правая панель показывает график воспроизводимости для сайтов связывания ERα с высоким уровнем сигнала ChIA-PET (50 или более PET на сайт).

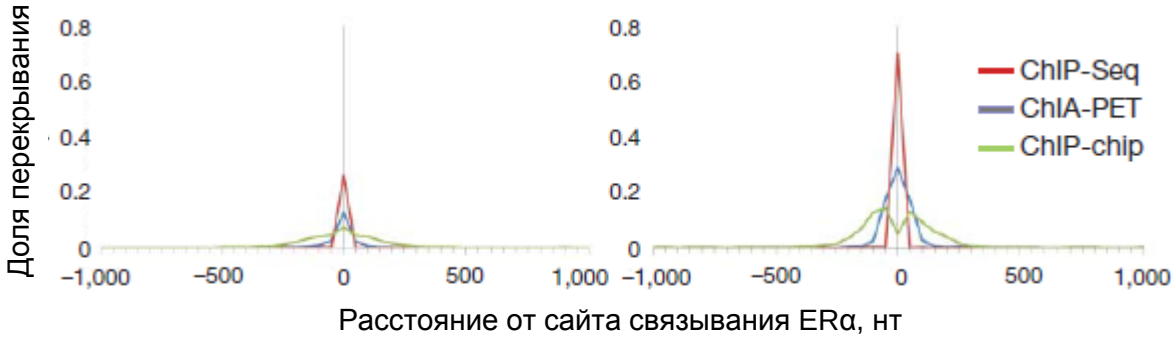


Рис. 5.6. Анализ воспроизводимости экспериментов иммунопреципитации связывания ERα в геноме человека по перекрыванию с данными ChIA-PET, данными ChIP-seq и данными ChIP-chip экспериментов [21].

Из рисунка видно четкое совпадение локализации сигнала по определению сайтов связывания ERα по данным ChIA-PET и ChIP-seq (в пределах 100 нуклеотидов). Высокий уровень сигнала ChIA-PET имеет более высокую долю пересечения, а значит лучшую воспроизводимость по другим экспериментам. Сайты связывания ERα, определенные с помощью ChIP-chip экспериментов, реже встречаются в сайтах ChIA-PET, и расстояние между ними чуть выше, что связано с особенностями технологии ChIP-chip.

Проведен компьютерный анализ распределения хромосомных контактов относительно генов и участков модификаций хроматина. Показана ассоциация контактирующих сайтов связывания ERα с комплексом полимеразы II и модификацией гистонов H3K4me3 (рис. 5.7). Видно обогащение сигнала для сайтов ERα, выявленных с помощью метода ChIA-PET.

Из рисунка 5.7 видно более сильную ассоциацию (повышенные профили в центре сайта) для всех маркеров для сильных взаимодействий ChIA-PET и уменьшение ассоциации для умеренных взаимодействий и неконтактирующих сайтов, найденных ChIA-PET.

Действительно, присутствие маркера открытого хроматина H3K4me3 для ССТФ ERα, определенных с помощью ChIP-seq, может использоваться для предсказания сайтов в геноме, также как и присутствие сигнала связывания ТФ FOXA1, как было представлено в предыдущем разделе.

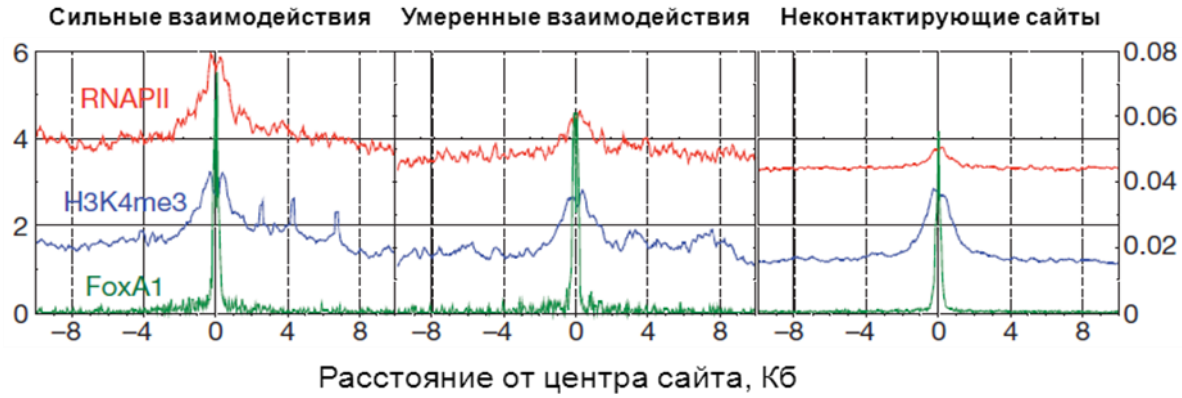


Рис. 5.7. Ассоциация сайтов связывания ERα, определенных с помощью ChIA-PET (Fullwood et al., 2009) различной интенсивности (три панели слева направо – от сильных взаимодействий до отсутствия контактов) с ChIP-seq профилями РНКполимеразы II, модификацией хроматина H3K4me3 (по данным ChIP-seq, левая ось Y) и ССТФ FOXA1 (данные ChIP-on-chip, правая ось Y).

Пример экспериментального подтверждения взаимодействий (контактирующие сайты помечены цифрами) через 3C эксперимент (на отдельном чипе) показан на рисунке 5.8.

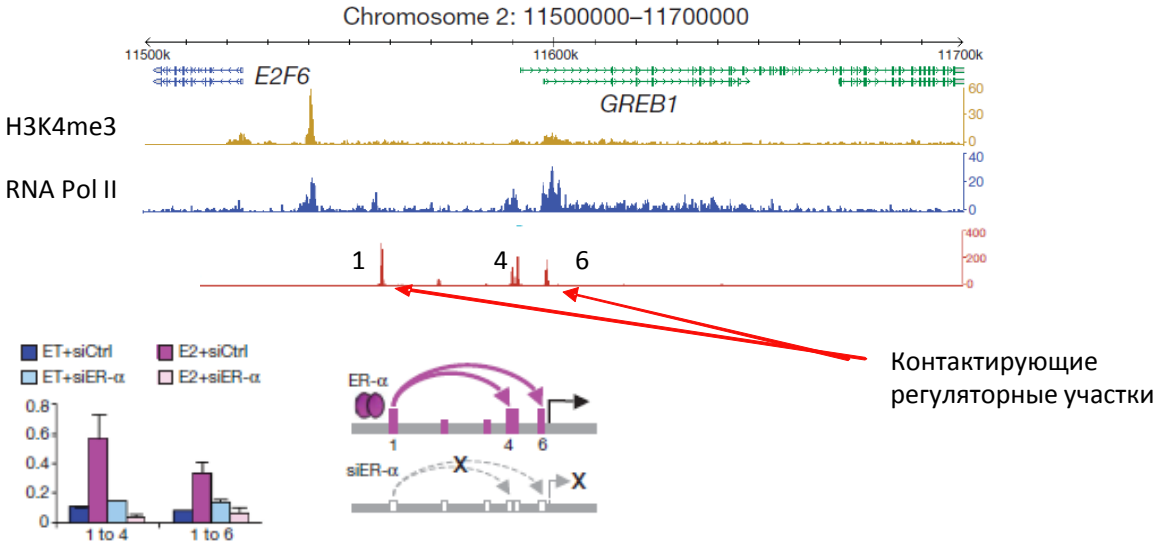


Рис. 5.8. Контакты между регуляторными участками, обусловленными связыванием ERα в гене *GREB1*, и экспериментальная проверка методом 3C при нокауте транскрипта рецептора эстрогенов посредством РНК-интерференции [21].

С помощью экспериментальной проверки 3C были подтверждены контакты между регуляторными участками, обусловленные связыванием ERα в гене *GREB1*. Из рисунка (нижняя панель) видно, что контакты между участками 1 и 4, и 1 и 6, появляются только при обработке эстрадиолом (E2) и функционирующем ERα. При

нокауте ERα с помощью РНК-интерференции контакты не образуются, значит, они вызваны именно действием димера ERα [21].

Следующий рисунок 5.9 представляет обобщение полученных представлений, общую схему структуры петель хромосом в ядре, которые образуются из-за локальных контактов, вызываемых действием белковых комплексов.

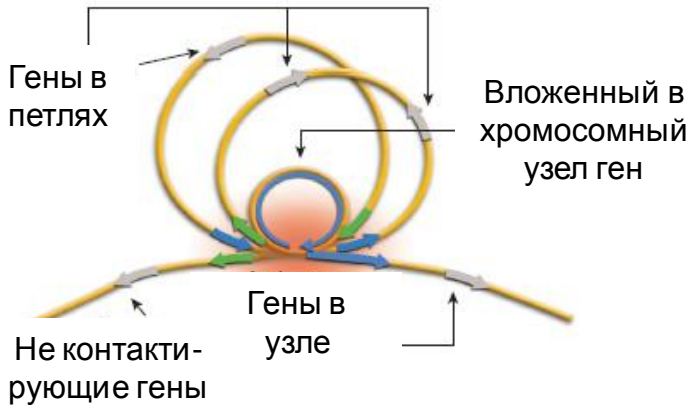


Рис. 5.9. Общая схема структуры петель хромосом и расположения генов, определяемая методом ChIA-PET [21]. Представлены контактирующие участки хромосом (узлы), петли, содержащие гены.

Схематически представлены контактирующие участки хромосом – петли и домены, определяемые методом ChIA-PET, и относительное расположение генов.

5.4. Хромосомные контакты, опосредованные комплексом РНК-полимеразы II в геноме человека

С помощью метода ChIA-PET исследовались полногеномные данные о сайтах связывания транскрипционных факторов и хромосомных контактах, опосредованных комплексом РНК-полимеразы II [12]. Задача стоит шире, чем исследование отдельного транскрипционного фактора и его сайтов связывания в геноме.

Фундаментальный вопрос биологии состоит в том, как гены и их регуляторные районы структурно организованы для регуляции транскрипции. Механизм пространственной координации транскрипции в клетках эукариот оставался неясным, несмотря на отдельные эксперименты микроскопии. Наблюдение флюоресценции in situ дало основания предполагать, что активная транскрипция не распределена равномерно в ядре клетки, но сконцентрирована дискретно в больших пространственных участках в ядре клеток млекопитающих, благодря чему можно предположить пространственную сближенность районов транскрипции в так называемых «транскрипционных фабриках» [503].

Оставался открытым вопрос, когда участки хромосом, содержащие транскрибируемые гены, комплекс РНК-полимеразы II (RNAPII) и другие белковые компоненты транскрипционной машины оказываются физически сближены для координации транскрипции генов в клетках млекопитающих. Был необходим глобальный метод высокого разрешения для описания функциональных взаимодействий хроматина относительно регуляции транскрипции. Для глобального исследования всех активных промоторов и соответствующих регуляторных районов генов и понимания, как они организованы и ассоциированы друг с другом in vivo, был применен метод ChIA-PET и проанализированы взаимодействия хроматина, ассоциированные с комплексом РНК-полимеразы II.

Результаты исследования дают представление о трехмерной организации и взаимодействии активных промоторов генов и регуляторных районов и представляют архитектурную модель, в которой соотносящиеся гены на мегабазном расстоянии в хромосомах со-организованы для эффективной и, возможно, кооперативной транскрипции.

Были проанализированы 5 различных человеческих клеточных линий (MCF7, K562, HeLa, HCT116 и NB4), используя метод ChIA-PET с антителом к белкам комплекса РНК-полимеразы II (8WG16), распознающим CTD домен (белок большой субъединицы комплекса для широкого набора видов млекопитающих). Используемые клеточные линии происходят от разных тканей, достаточно широко представляя клетки организма человека.

В пилотном исследовании около 20 миллионов уникально картированных парных концов было сгенерировано для каждого ChIA-PET эксперимента, что дало два полногеномных набора данных: геномное расположение сайтов связывания РНК-полимеразы II после иммунопреципитации (ChIP-seq эксперимент) и парные положения контактирующих удаленных участков хромосом, опосредованные полимеразным комплексом. Были получены данные как о внутрихромосомных, так и о межхромосомных взаимодействиях, при этом подавляющее большинство контактов хроматина, идентифицированных ChIA-PET, были внутрихромосомными. Двадцать пять внутрихромосомных и семь межхромосомных взаимодействий были дополнительно проверены экспериментально одним из методов 3C, DNA-FISH или двумя методами вместе (Li et al., 2012).

Пример экспериментально подтвержденных контактов в районе гена *KLF4* представлен на рисунке 5.10.

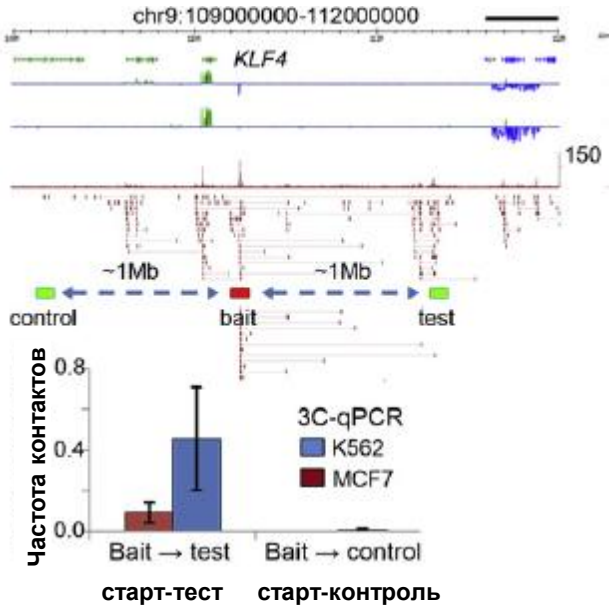


Рис. 5.10. Пример экспериментально подтвержденных контактов хроматина, определенных с помощью ChIA-PET, в районе гена *KLF4* (Fullwood et al., 2009).

Для исследования данных ассоциированного с РНК-полимеразой II (RNAPII) человека интерактома хроматина (набора всех взаимодействий хроматина), прочтения секвенированных фрагментов ДНК экспериментов ChIA-PET из 6 пилотных экспериментов для культур клеток человека были скомбинированы вместе для анализа (Таблица 5.2).

Таблица 5.2

Пилотные эксперименты ChIA-PET (тип клеток)	Число уникальных PET	Пики RNAPII	Кластеры взаимодействий
HCT116	8,749,703	17,394	19,264
K562	23,188,484	13,112	17,686
HeLa	19,079,666	15,333	200,952
NB4	14,023,893	12,707	54,232
MCF7 (реплика 1)	38,356,322	10,370	12,626
MCF7 (реплика 2)	22,967,674	11,234	8,111
Общее число	118,523,881	14,604	892,991

Используя внедренный контрольный бар-код и статистический анализ, было оценено качество данных, отфильтрован технический шум сигнала секвенирования, и определены надежные, с высоким уровнем значимости, сайты связывания и кластеры парных концов (PET), соответствующие хромосомным контактам. Из комбинированного набора данных было выделено 14,604 сайтов связывания РНК-

полимеразы II с высоким уровнем значимости (FDR<0.05) и 19,856 внутрихромосомных контактирующих PET кластеров.

Построена статистика распределения участков связывания РНК-полимеразы II в геноме относительно генов (на основе аннотации генов в базах данных RefSeq, UCSC genes). Профиль связывания РНК-полимеразы II относительно генов в культурах клеток человека (рис. 5.11) построен по участкам связывания в геноме человека, определенных методом ChIA-PET. Такое распределение указывает на возможность хромосомных контактов между этими участками, образования петли между 5’ и 3’ районами гена, что способствует переходу РНК-полимеразы II после окончания транскрипции вновь в промоторный участок для нового цикла транскрипции.

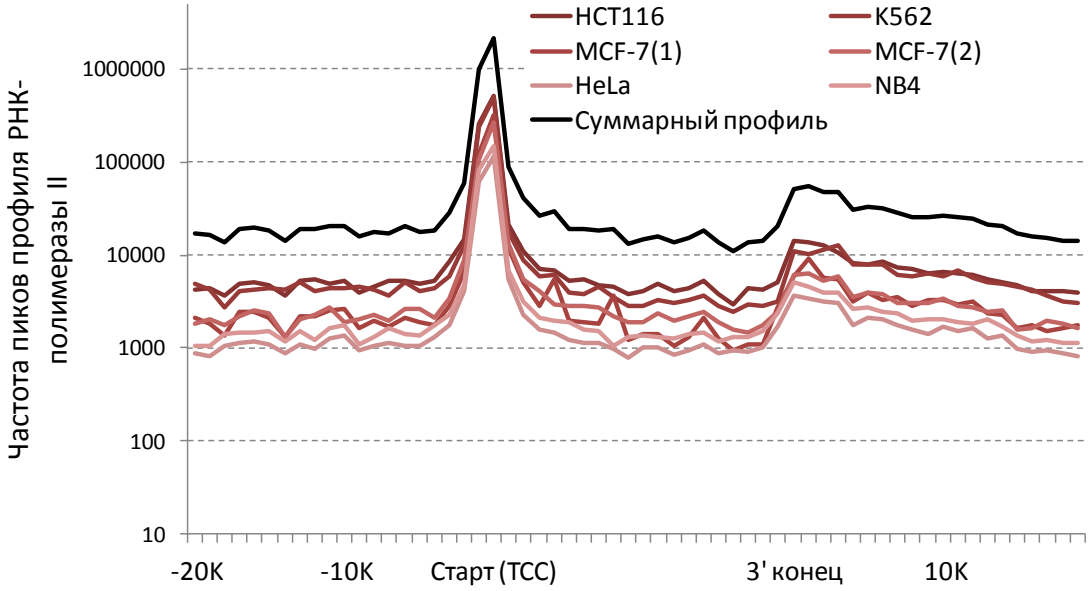


Рис. 5.11. Профиль связывания РНК-полимеразы II относительно генов в культурах клеток человека [12].

Большинство (83%) сайтов связывания РНК-полимеразы II (RNAPII) в комбинированном наборе данных расположены проксимально к 5’ точке старта транскрипции генов (TCC, или TSS - Transcription Start Sites) (Рисунок 5.11). Наблюдается также четкий, но относительно более слабый пик связывания в 3’ районе генов (обозначается как TES - Transcription End Sites). Такой же паттерн распределения сайтов связывания вдоль гена присутствует во всех библиотеках секвенирования для разных типов клеток.

Из общего числа сайтов связывания РНК-полимеразы II 9487 (65%) вовлечены в удаленные взаимодействия хроматина, и эти сайты имеют более высокий сигнал связывания РНК-полимеразы II, чем сайты, не вовлеченные во взаимодействия, что указывает на большую активность РНК-полимеразы II в петлях хроматина, в контактирующих конформациях. На рисунке представлена объемная гистограмма

распределения пиков связывания РНК-полимеразы II для двуз типов участков - без удаленных взаимодействий, и со взаимодействиями.

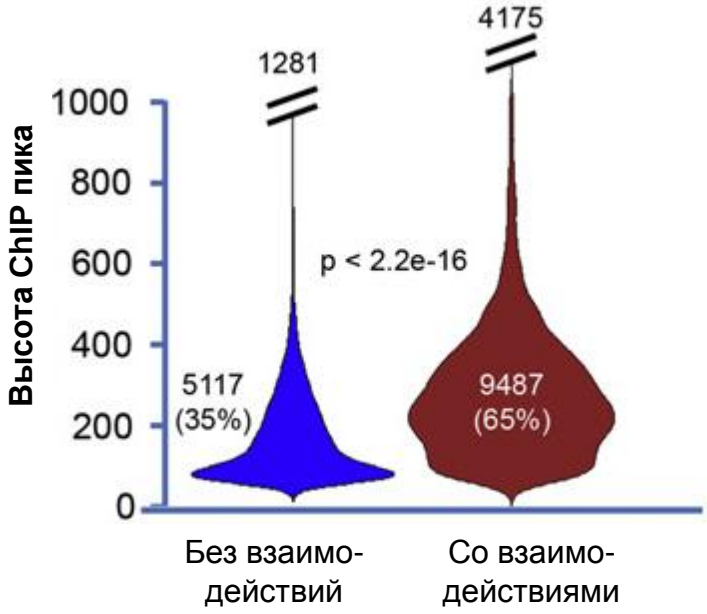


Рис. 5.12. Распределения интенсивности связывания РНК-полимеразы II (высота пика) по методу ChIA-PET для участков без удаленных взаимодействий и со взаимодействиями [12].

Были идентифицированы три основных типа взаимодействий вокруг промоторов генов в комбинированном пилотном наборе данных: внутри-генные (контакт - промотор к внутренним районам гена, 938 контакта, или 5%), вне-генные (промотор к дистальным регуляторным элементам, таким как энхансеры, 6530, или 33%) и межгенные (промотор к промоторам различных генов, 8282, или 42%) взаимодействия. Также отмечается дополнительная категория хромосомных контактов, которая включает промежуточные предполагаемые взаимодействия энхансер-энхансер (4106, или 20%). Некоторые взаимодействия (2341, или 12%) были представлены отдельными дуплексными (двойными) взаимодействиями между двумя взаимодействующими районами контактов, в то время как большинство (17515, или 88%) далее агрегируются в комплексы взаимодействий (1544 комплексных взаимодействий, включающие три и более района).

Можно выдвинуть гипотезу, что одиночное связывание РНК-полимеразы II в промоторных районах генов, без вовлечения во взаимодействия, отражает базальную функцию промотора для транскрипции генов. Поэтому такие взаимодействия можно назвать промоторной, или базальной промоторной моделью (Basal Promoter - BP, в дальнейших обозначениях). Напротив, ассоциированные с РНК-полимеразой II взаимодействия хроматина могут представлять собой структурную основу для комплексных регуляторных механизмов. Эти основные взаимодействия хроматина

могут далее агрегировать в более сложные архитектуры, которые классифицируются как регуляторная одногенная модель (single-gene - SG) и мультигенная модель (MG),соответственно, в зависимости от числа вовлеченных генов в модели (Рисунок 5.13). Мультигенная (MG) структурная модель включает меженные промотор-промоторные взаимодействия и может также включать внутри-генные и вне-генные взаимодействия энхансер-промотор. В то же время одиночная модель (SG) может включать одно или несколько взаимодействий энхансера с одним промотором. Несколько мультигенных комплексов взаимодействия, дистально расположенных на хромосоме, и даже на разных хромосомах, далее могут формировать мультигенные комплексы взаимодействий высокого порядка.

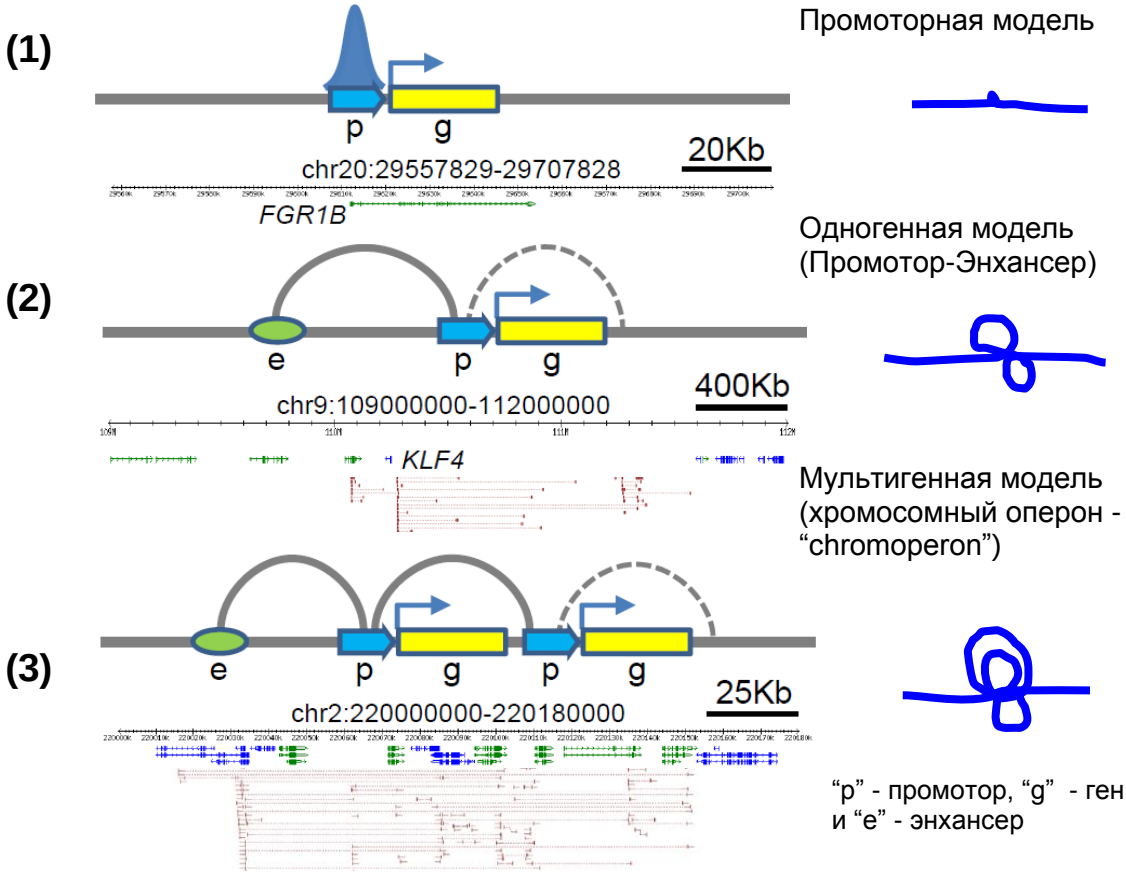


Рис. 5.13. ChIA-PET. Модели промоторных, энхасерных и мультигенных контактов, опосредованных комплексом РНК-полимеразы II [12].

Большая часть моделей контактов хроматина имеет размер расположения на хромосоме в 150-200Кб, а некоторые комплексы контактов захватывают несколько мегабаз. Хотя общее число мультигенных комплексов в комбинированном наборе данных составляет чуть более тысячи (1328 MG комплексов), число генов в этих комплексах на порядок больше - 11723 генов (Рис. 5.14), указывая на то, что промотор-

промоторные взаимодействия широко распространены и могут играть доминирующую роль в транскрипционной регуляции экспрессии генов.

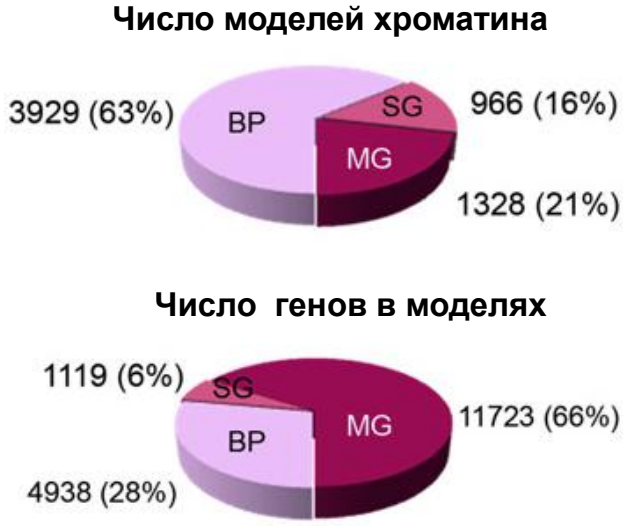


Рис. 5.14. Число моделей контактов хроматина (промоторной, одногенной и мультигенной - BP, SG и MG) и число генов в геноме человека, входящих в эти модели [12].

Рассмотрим отличительные геномные особенности моделей контактов, ассоциированных с РНК-полимеразы II (BP, SG и MG), и механизмы регуляции транскрипции в геноме человека. С помощью компьютерного анализа данных ChIA-PET были охарактеризованы структурные черты (геномные свойства), функциональные особенности (транскрипционная активность генов), и эпигеномные маркеры (состояние хроматина).

Отличия геномных характеристик одно- и мульти-генной моделей хромосомных контактов

Для определения геномных характеристик структур взаимодействий хроматина, ассоциированных с РНК-полимеразой II, с координатами таких хромосомных контактов были соотнесены несколько геномных дескрипторов (аннотационных карт), связанных с экспрессией генов в геноме человека [139], включая GC состав, плотность расположения генов, присутствие SINE/LINE повторов, длину гена, и отношение длин интрон/экзон. В таком анализе, мультигенные комплексы взаимодействий (MG) были значительно обогащены высоким GC-составом, более высокой плотностью генов и SINE повторов, и более низкой плотностью LINE повторов в сравнении с комплексами взаимодействий одиночных генов и комплексов районов базального промотора, предполагая, что мультигенные комплексы расположены в районах активного хроматина, районах генома с активной транскрипцией.

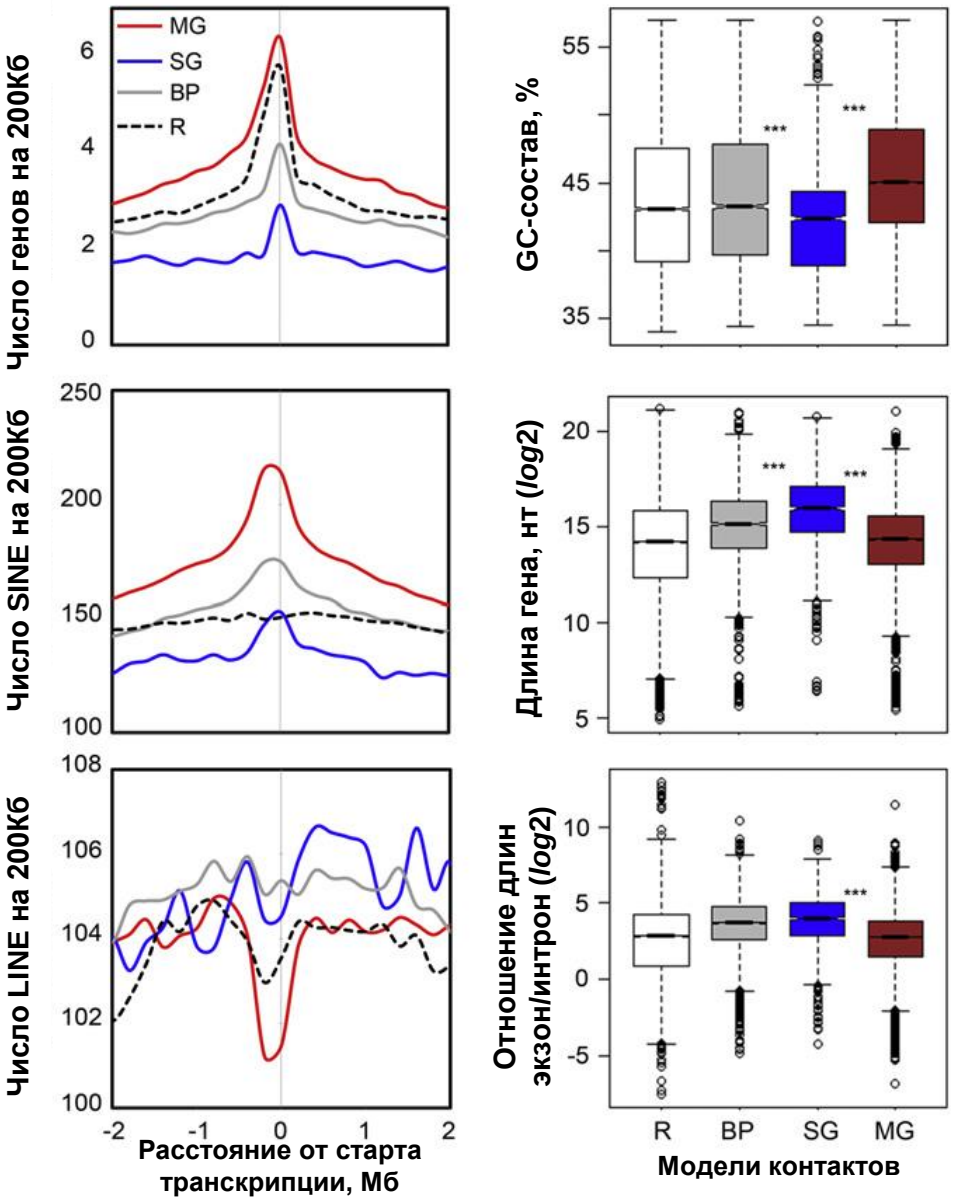


Рис. 5.15. Модели контактов хроматина и геномные характеристики [12].
(Левая панель) Профиль связи плотности расположения генов, повторяющихся последовательностей SINE и LINE вокруг стартов транскрипции генов для различных моделей хромосомных контактов. Использовались старты транскрипции для аннотации генов RefSeq. Красная линия обозначает мультигенную модель контактов (MG), синяя - одногенную модель (SG), серая - модель базального промотора (BP), и серная прерывистая линия означает остальные гены, не участвующие в контактах (R).
(Правая панель) Бокс-схема показывает распределение GC состава в GC изохорах вокруг различных моделей контактов, дину гена, и отношение длин интрон/экзон для генов RefSeq, вовлеченных в эти модели. Тройные звездочки (***) означают уровень значимости p<2.2E-16. Красный цвет показывает распределение для мультигенной модели (MG), синий для одногенной (SG), серый и белый цвета - для модели базального промотора и остальных генов.

Кроме того, гены в мультигенных комплексах были относительно короче по длине, чем гены других рассмотренных категорий, что является еще одним свойством высокоэкспрессирующихся генов [140]. Напротив, геномные локусы, ассоциированные

с комплексами хромосомных контактов одиночных генов (SG), лежат в районах с пониженной плотностью генов и SINE элементов.

Геномные свойства промотор-ориентированных моделей хроматина

Гены, вовлеченные в одногенные комплексы контактов (SG), значительно длиннее и имеют большее отношение длин интрон/экзон, чем гены других моделей контактов хроматина. Эти наблюдения подтверждают, что гены с энхансер-промоторными взаимодействиями в одногенных комплексах скорее являются тканеспецифичными или генами, регулируемыми во время развития организма. Полученный результат соответствует ранее показанным данным о том, что гены в бедных генами районах генома имеют тенденцию к большей длине и имеют более высокое соотношение некодирующей части к кодирующей, чем гены домашнего хозяйства [140, 141].

Контактирующие гены имеют коррелированную экспрессию

Для исследования транскрипционной активности и функции генов, вовлеченных в различные модели контактов хроматина, анализ был сконцентрирован на клеточной линии MCF-7, поскольку это хорошо характеризованная линия раковых клеток человека с существующими взаимно дополняющими наборами данных, включая RNA-Seq, данные GRO-Seq [286], наборы экспрессионных данных на микрочипах, в том числе временные (time-course) серии [21].

В соответствии с комбинированным пилотным набором данных, наибольшая часть (90%) сайтов связывания РНК-полимеразы в клетках MCF7 расположена проксимально к известным промоторам генов, и большинство (97%) генов с детектированным присутствием РНК-полимеразы II, на их промоторах имели определенную транскрипционную активность, детектированную методом RNA-seq. Контактирующие сайты связывания РНК-полимеразы II, которые были дистально расположены по отношению к промоторам, включали внутригенные и внегенные (внешние по отношению к генам) регуляторные элементы, такие как энхансеры. Приблизительно 45% внешних по отношению к генам дистальных регуляторных элементов имели выраженный, детектируемый сигнал присутствия РНК, который может отражать присутствие некодирующих РНК (нкРНК) транскриптов.

10108 (90.6%) пиков связывания РНК-полимеразы II находились в геноме в проксимальных промоторах, и подавляющее большинство, 9779 (96.8%), из соответствующих генов экспрессировались.

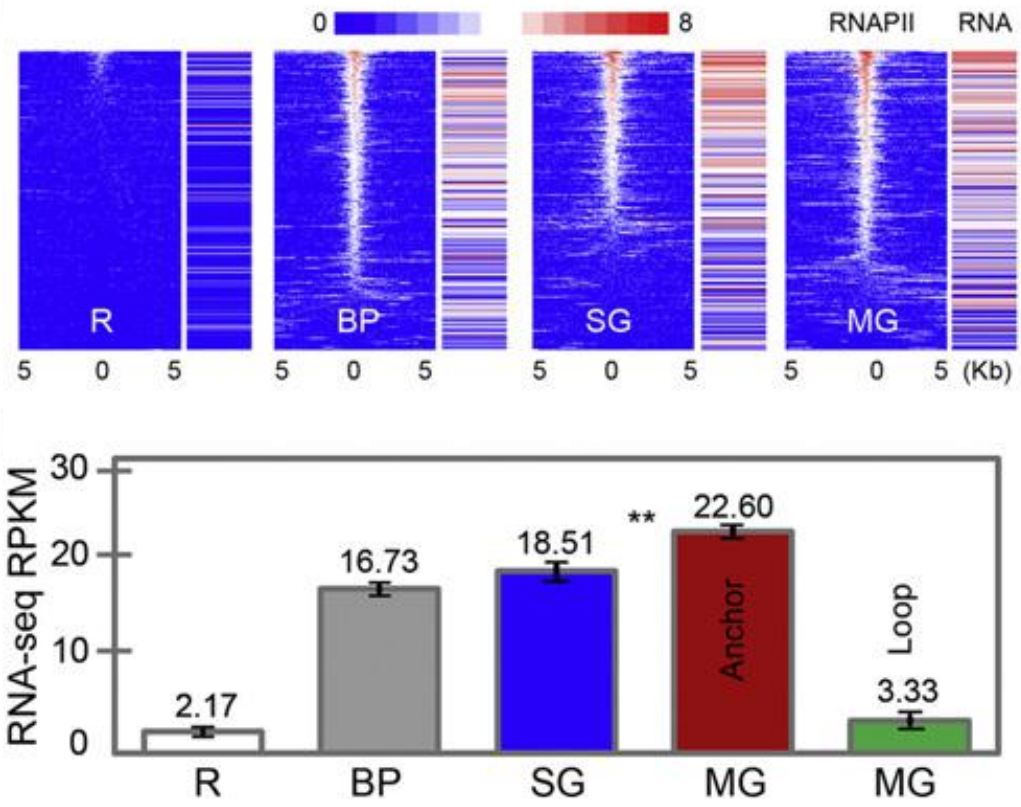


Рис. 5.16. Транскрипционная активность ассоциированных с РНК-полимеразой II моделях хроматина в клетках MCF-7 [12].
(Верхняя панель) Корреляция связывания РНК-полимеразы II в моделях базального промотора (BP), одиночного гена (SG) и мультигенной модели (MG) с уровнем транскрипции генов, измеренным с помощью RNA-seq. Тепловая карта сигнала РНК-полимеразы II показывает интенсивность связывания вокруг стартов транскрипции генов (±5 Кб) вместе с соответствующей интенсивностью транскрипции гена.
(Нижняя панель) Гистограмма уровней экспрессии генов в трех моделях хромосомных контактов (BP, SG, и MG). Показаны средние значения уровня экспрессии генов, измеренные с помощью RNA-seq (в единицах RPKM) и стандартная ошибка среднего для генов из соответствующих моделей. Комплексы мультигенной модели MG также содержат «якорные гены» (anchor) (проксимальные к контактирующим участкам контактов - «якорям») и «петлевые гены» (loop) (удаленные от участков контактов, находящиеся в петлях между «якорями»). Оставшиеся гены (R), не связанные РНК-полимеразой II в эксперименте, были включены в качестве контроля. Двойная звездочка (**) указывает значительную разницу между средними значениями экспрессии генов из моделей SG и MG (уровень значимости $p < 4.02E-08$).

Для генов, ассоциированных с тремя представленными моделями организации контактов хроматина, был проанализирован уровень транскрипции по числу прочтений RNA-Seq. Как показано на рисунке, в целом связывание РНК-полимеразы II на промоторах хорошо коррелирует с уровнем экспрессии соответствующих генов. Интересно отметить, что гены, вовлеченные в одногенную и мультигенную модели, показали большую корреляцию между уровнем связывания РНК-полимеразы II и сигналом RNA-seq: линейный коэффициент корреляции Пирсона (PCC) равен 0.46 и 0.45 соответственно), выше по сравнению с генами модели базального промотора (PCC

составил 0.24). Кроме того, гены, вовлеченные в комплексные взаимодействия хроматина, особенно в мультигенных комплексах, имели значительно более высокие уровни экспрессии, чем гены модели базального промотора. Примечательно, что гены, расположенные в промежутках петель хроматина, в целом имеют пониженную транскрипционную активность. Эти экспериментальные данные показывают, что промотор-промоторные взаимодействия в мультигенных комплексах ассоциированы с повышенной транскрипционной активностью, что соответствует присутствию ассоциированных геномных характеристик.

На следующем шаге были проанализированы паттерны экспрессии генов, присутствующих в контактирующих районах, используя данные экспрессионных микрочипов, полученные из 84 тканей человека, в базе данных BioGPS [147]. Было найдено различие в представленности тканеспецифичных генов и генов домашнего хозяйства в трех рассмотренных моделях контактов хроматина. Число различных тканей и типов клеток из 84, в которых детектируется экспресссия гена, использовалась как цифровая оценка тканеспецифичности - чем меньше это число, тем более тканеспецифична экспрессия рассматриваемого гена. Чем больше это число, чем шире экспрессия, тем менее

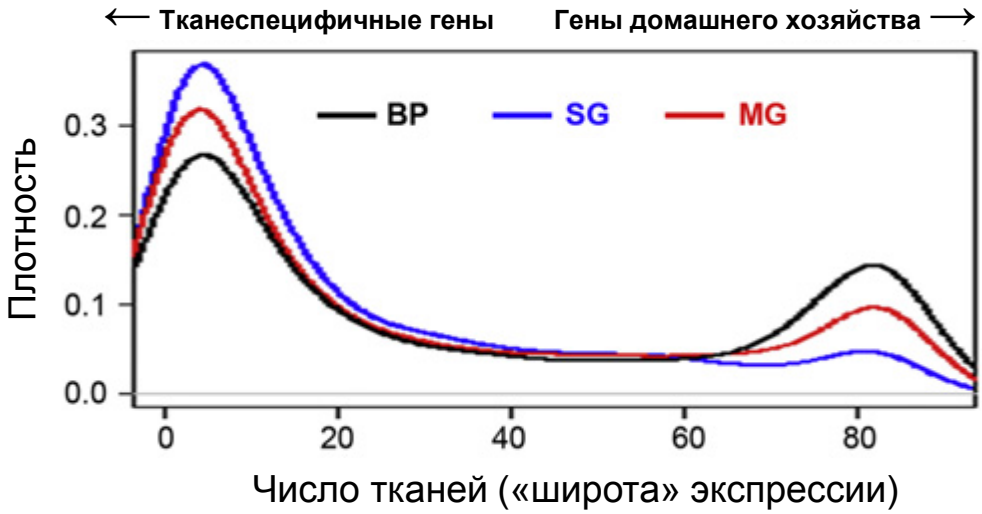


Рис. 5.17. Тканеспецифичность экспрессии (число тканей организма человека, в которых экспрессируется ген) для генов, представленных в трех моделях хроматиновых контактов [12].

Уровень значимости (*P*-value) для различия распределений, показанных на рисунке, был рассчитан с помощью непараметрического теста Краскала-Уоллиса. Наибольшее число генов в одногенных комплексах контактов с энхансер-промоторными взаимодействиями - тканеспецифичны, в соответствии с более ранними доказательствами того, что уровни экспрессии генов развития и тканеспецифичных

генов сильно модулируются через цис-регуляторные взаимодействия и транс-действующие белковые факторы транскрипции [517-519], и в соответствии с их геномными характеристиками (меньшая плотность генов, большая длина гена и более высокое соотношение интрон/экзон). Напротив, гены, вовлеченные в мультигенные комплексы контактов, также как и гены промоторных моделей, характеризуются категориями как тканеспецифичных генов, так и генов домашнего хозяйства. Таким образом, функции базального промотора и промотор-промоторные взаимодействия применимы для тканеспецифичных функций и функций генов домашнего хозяйства, требующих постоянной и стабильной экспрессии. Эти наблюдения подтверждаются нормализованной оценкой CpG состава и отклонением GC в промоторных районах генов этих моделей.

Поскольку промотор-промоторные взаимодействия собирают вместе в пространстве ядра множество различных генов, они могут обеспечить идеальную топологическую структуру для потенциальной транскрипционной координации обеих групп генов, как тканеспецифичных, так и домашнего хозяйства. Действительно, это наблюдение согласовано с существованием геномных доменов, «перевалов», или «водоразделов» (“ridges”), являющихся доменами генов с высокой транскрипцией, которые содержат как гены домашнего хозяйства, так и тканеспецифичные гены [139]. Поскольку большое число генов находится именно в мультигенных комплексах контактов, можно предположить, что промотор-промоторные взаимодействия служат доминантным механизмом для транскрипционной регуляции как генов домашнего хозяйства, так и тканеспецифичных генов в геномах млекопитающих.

Далее рассматривался вопрос о совместной транскрипционной координации генов с промотор-промоторными взаимодействиями. Данные RNA-seq показывают, что большая часть пар генов с промотор-промоторными взаимодействиями экспрессируются вместе на высоком уровне. Чтобы далее оценить координированную транскрипцию пар генов в различных условиях, был выполнен анализ коэффициента корреляции Пирсона между уровнями экспрессии генов, используя временную серию данных GRO-Seq после индукции эстрадиолом [286], когда измерялся уровень транскрипции генов.

Распределение значений линейного коэффициента корреляции Пирсона (PCC) между уровнями экспрессии контактирующих генов для пар генов, контакты которых опосредованы РНК-полимеразой и связыванием транскрипционногоф атора ERα представлены на рисунке. В качестве контроля использовалисьслучайные пары генов,

взятые из геномных локусов такого же размера, как и гены в участках хромосомных контактов.

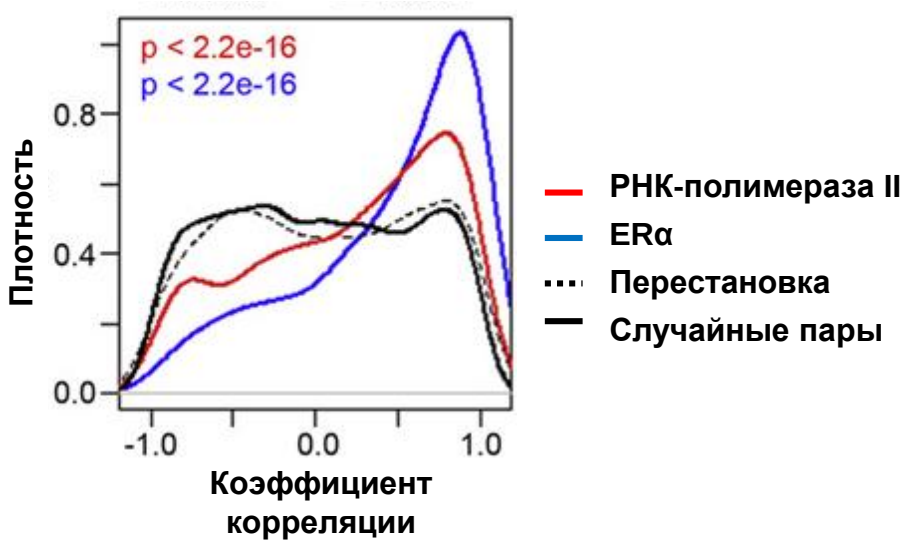


Рис. 5.18. Распределение значений линейного коэффициента корреляции Пирсона (PCC) между уровнями экспрессии контактирующих генов для пар генов, контакты которых опосредованы РНК-полимеразой II (красная линия), связыванием рецептора эстрогенов генов ERα (синяя линия), для контрольных «переставленных» пар генов из разных контактов (пунктир), и из случайно выбранных пар генов в геноме из районов, из районов, имеющих тот же размер и ту же плотность генов, как районы мультигенных комплексов (черная сплошная линия).

Наблюдались значимые коэффициенты корреляции в таких парах (уровень значимости p < 2.2E-16). Интересно отметить, что корреляция была даже выше для пар генов, контакты в которых опосредованы ERα, - пар, полученных в работе автора [21], что предполагает более сильную корреляцию транскрипции для генов, вовлеченных в мультигенные комплексы, опосредованные специфическими факторами транскрипции.

Подобная корреляция наблюдалась также на других наборах данных генной экспрессии. Как ожидалось, гены домашнего хозяйства и гены, принадлежащие к одному классу генных онтологий (ГО), показывали даже большую корреляцию, чем остальные гены. В целом проведенный анализ показал, что значительная доля пар генов, имеющих промотор-промоторные взаимодействия, имеет тенденцию к кооперативной транскрипции.

Мультигенные комплексы задают структурные основы для транскрипционной ко-регуляции

Коррелированная экспрессия взаимодействующих генов позволяет предположить, что мультигенные комплексы контактов могут обеспечивать молекулярную основу для постулируемой «транскрипционной фабрики», в которой транскрипты и компоненты комплекса РНК-полимеразы сконцентрированы в дискретных объемных точках (фокусах) в ядре, что наблюдается с помощью меченых

in situ транскриптов и в экспериментах по иммунофлюоресцентному окрашиванию [503].

Чтобы оценить связь между мультигенными комплексами, выявленными с помощью ChIA-PET, и транскрипционными фабриками, был выполнен эксперимент трехмерной флюоресцентной гибридизации in situ - 3D DNA-FISH, используя пробы, представляющие различные мультигенные комплексы в комбинации с иммуноферментативным окрашиванием РНК-полимеразы II (RNAPII-IF) в ядрах клеток линии MCF-7. Все эксперименты на четырех случайно выбранных локусах выявили значимую ассоциацию локусов мультигенных комплексов с пространственным расположением (фокусом) комплекса РНК-полимеразы II. Такая проверка дала дальнейшее доказательство представления о мультигенном комплексе хромосомных контактов как структурообразующей схемы кооперативной транскрипции генов.

Более того, ряд генных семейств значимо перепредставлен в мультигенных комплексах (уровень значимости $p < 0.006$), таких как *HIST, ZNF, KRT, HOXC*, и так далее. Рассматривая семейство *HIST1H* в качестве примера, были определены 58 генов этого семейства на хромосоме 6, формирующих три мультигенных комплекса, эти три комплекса в более широком определении могут рассматриваться как суперкомплекс контактов более высокого порядка, предполагая, что все гены *HIST1H* организованы в единую архитектуру хроматина для координированной транскрипции. Все гены *HIST1H* активно транскрибируются в клетках линий MCF7 и K562, и строго совместно регулированы в различных тканях и условиях культивирования клеток. Интересно отметить, что ген *HFE*, который не является частью семейства *HIST1H*, но расположен в середине первого мультигенного комплекса *HIST1H*, не привязан к контактирующим участкам и не экспрессируется. Подобно этому, гены, расположенные в районах промежуточных петель между тремя взаимодействующими комплексами *HIST1H,* имеют относительно меньшую транскрипционную активность и намного меньше координированы для ко-регуляции в различных тканях и клеточных состояниях. Этот пример дает представление, о том, как мультигенные комплексы могут организовать гены с подобными функциями совместно для координированной экспрессии.

Связь хромосомных контактов и модификаций хроматина (метилирование и ацетилирование) гистонов (гистона H3) исследовалась статистически в масштабе генома. На рисунке показаны профили модификаций хроматина для участка хромосомы 2 человека, содержащий ген *PCBP1*.

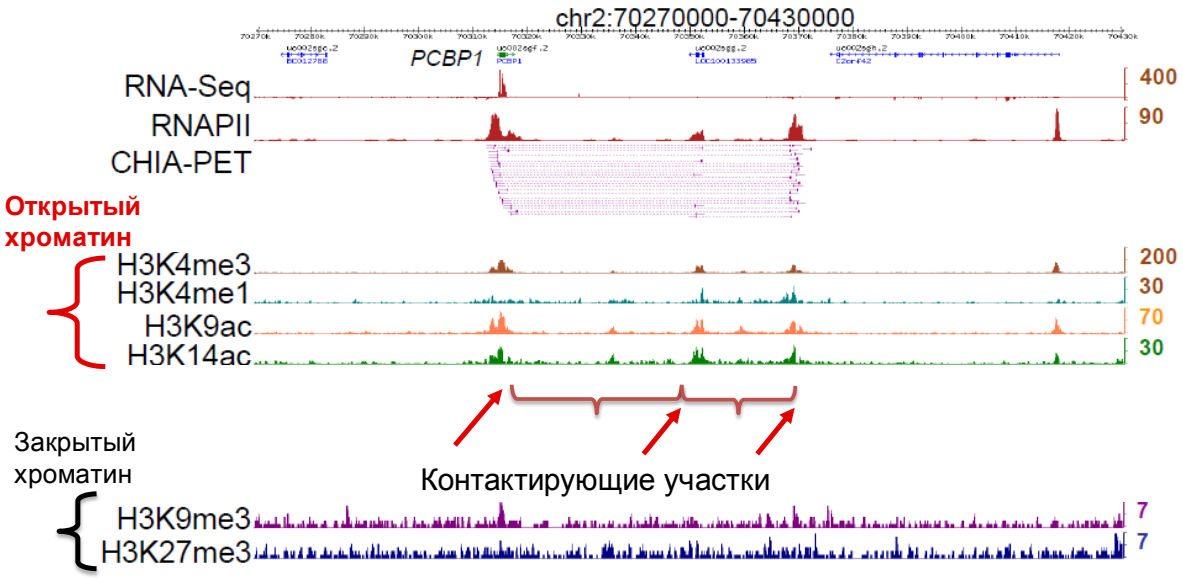


Рис. 5.19. Пример профилей ChIP-seq модификаций гистонов, соответствующих открытому и закрытому состоянию хроматина (группы показаны слева), в участках хромосомных контактов, найденных с помощью ChIA-PET (отмечено внизу) на хромосоме 2 человека [12].

Хроматин в контактирующих участках открыт – гистоны имеют маркеры модификации активной транскрипции – H3K4me3, H3K9ac (видны пики профиля) в то же время модификации репрессии транскрипции не имеют пиков (равномерный шум).

Важнейшее значение имеют модификации хроматина, прежде всего гистона H3, модификации лизина в позициях 4, 14, 36, включающие метилирование и ацетилирование, связанные с доступностью ДНК для связывания белковых факторов транскрипции [19]. Ассоциации модификаций хроматина с активацией работы генов через изменение структуры хроматина в местах связывания транскрипционных факторов позволяют предсказать сайты связывания в полногеномной шкале, что было показано детально для связывания рецептора эстрогена ERα [13]. Более того, исследованные ассоциации хромосомных контактов, опосредованных белком ERα, также связаны с активацией хроматина, в частности модификациями H3K4me3, H3K4me1 для генома человека методом ChIA-PET. Исследование контактов хромосом, опосредованных комплексом РНК-полимеразы II в культурах клеток человека, также подтвердило ассоциацию с сайтами связывания факторов транскрипции и модификациями хроматина [12].

Функциональный тест синергичности мультигенных взаимодействий

Для дальнейшего исследования предположения о том, что взаимодействующие гены в мультигенных комплексах могут определять структурную основу для ко-регуляции генной экспрессии, был выполнен набор экспериментов индукции

экспрессии и ее нарушения для проверки данного предположения. После сравнения наборов данных ChIA-PET для комплекса РНК-полимеразы II и ERα в клеточной линии MCF7, установлено, что мультигенный комплекс, опосредованный РНК полимеразой II, в локусе гена *GREB1* частично перекрывается в геноме с расположением петель хроматина, вызванных связыванием ERα, что позволяет предположить роль связывания ERα в формировании мультигенного комплекса. Был выполнен эксперимент с использованием siRNA по нокдауну (уменьшению уровня экспрессии) ERα в клетках MCF7, и наблюдению изменений взаимодействий хроматина и экспрессии генов в мультигенном комплексе *GREB1*. Несколько контактов хроматина в этом локусе были нарушены посредством трансфекции siERα, что было подтверждно экспериментами 3C. В дополнение к *GREB1*, который дает ответ увеличения экспрессии при индукции эстрогеном и уменьшает экспрессию при редукции (нокдауне) посредством siERα, наблюдается, что другие гены в этом комплексе - *E2F6*, *KCNF1* и *ATP6VC12* - также имеют различные уровни экспрессионного ответа на индукцию эстрогеном и редукцию посредством siERα нокдауна. Интересно отметить, что эти гены не контактировали непосредственно с ERα в их промоторных районах, но косвенно были ассоциированы с ERα через петли хроматина, образованные связыванием РНК-полимеразы II. Для сравнения, этот эффект не наблюдался в близлежащих генах, таких как *NOL10* и *HPCAL1*, которые находятся в других комплексах взаимодействий, сформированных RNAPII и не контактируют с ERα. Таким образом, эти результаты показывают, что специфичный стимул (эстроген) может вести к совместной активации генов организованных в мультигенный комплекс, сформированный РНК-полимеразой II, и эпигенетические изменения в одном локусе, в данном случае связывание ERα, могут менять транскрипционное состояние другого контактирующего гена. Функциональная значимость ко-регуляции таких независимых и неродственных генов требует дальнейшего изучения. Тем не менее, проксимально расположенные друг к другу гены действительно имеют тенденцию быть совместно регулированными независимо от функциональных различий между ними [184-186], и такие генные кластеры называют «нейтрально ко-экспрессирующимися кластерами» (“neutral co-expression clusters”) которые, как предполагают, появляются в результате эффекта нейтральной ко-эволюции [187].

Эпигенетические маркеры ассоциированы с районами контактов хроматина

Для исследования ассоциации транскрипционных факторов с хросомными контактами, опосредованными РНК-полимеразой II, были рассмотрены полногеномные

профили связывания 20 различных транскрипционных факторов в клетках K562, доступных из данных Консорциума ENCODE в аннотации Геномного браузера UCSC. Исследовалось обогащение числа прочтений ChIP-seq для этих факторов на участках каждой из трех моделей контактов хроматина, опосредованных РНК-полимеразой II, в наборе ChIA-PET данных в клеточной линии K562. Общие факторы транскрипции, такие как E2F4 и E2F6, непосредственно связываются в районах старта транскрипции генов и действуют, прежде всего, через промоторные районы (см. Figure 5B для конкретного примера). Напротив, специфичные транскрипционные факторы, такие как JunD и Max предпочтительно связываются в дистальных регуляторных районах и маркируют потенциальные энхансеры.

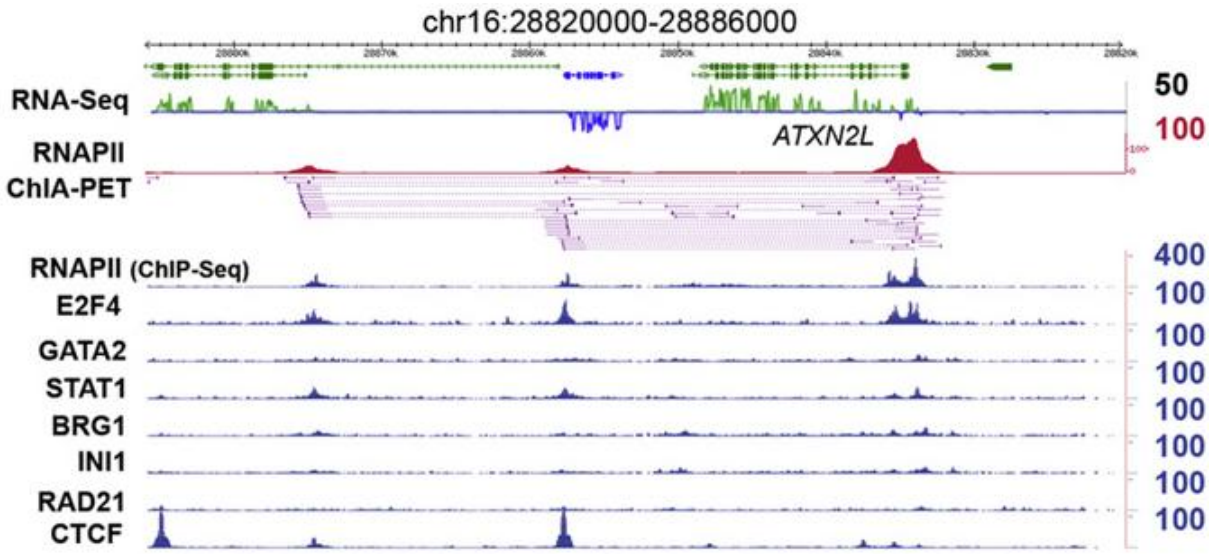


Рис. 5.20. Пример связи ChIA-PET контактов с сайтами связывания различных транскрипционных факторов (E2F4, GATA2, STAT1, BRG1, INI1, RAD21, CTCF) в геноме человека в районе гена *ATXN2L*.

Несколько факторов ремоделинга хроматина и белков, организующих структуру хроматина, таких как Ini1, Brg1 [191], CTCF и Rad21 [189] ассоциированы, прежде всего, с участками дистальными к стартам транскрипции, предполагая, что они могут опосредовать дальнедействующие взаимодействия с энхансерными районами. Гипотеза подтверждается независимыми данными о том, что Ini1 и Brg1, две субъединицы комплекса SWI/SNF, вовлечены в транскрипционные петли [190, 191]. Общее наблюдение для всех рассмотренных транскрипционных факторов состоит в том, что участки хромосомных контактов в мультигенном комплексе взаимодействий четко показывают повышенный уровень ChIP-seq сигнала связывания этих факторов. Такая ассоциация позволяет предположить, что кооперативное связывание факторов в ген-богатых доменах генома ведет к повышенной транскрипционной активности, и что эти транскрипционно активные домены открытого хроматина могут конвергировать в

отдельные специализированные транскрипционные фабрики, каждая из которых обогащена специфичным транскрипционным фактором или множеством общих факторов транскрипции.

Для сравнительного анализа использовались данные по модификациям хроматина, доступные в базе ENCODE. В целом, было установлено обогащение активных модификаций хроматина, маркирующих открытый, доступный для связывания хроматин и повышенную транскрипцию, вместе с недостатком репрессивных модификаций гистонов в сайтах взаимодействий РНК-полимеразы II, картированных по данным ChIA-PET в промоторных и дистальных регуляторных районах. Интересно отметить, что маркеры активного хроматина были самыми высокими в мультигенных комплексах, подтверждая, что именно такие комплексы являются транскрипционными узлами. Это наблюдение, совпадает с тем, что обогащение сигнала активных модификаций гистонов положительно коррелирует с присутствием РНК-полимеразы II (см. предыдущую главу настоящей работы - данные по связыванию ER).

Наблюдались похожие профили модификаций гистонов в клетках линии MCF-7, используя сгенерированные ранее данные [13].

Кроме того была использована дополнительная мера: отношение ChIP-seq в логарифмической шкале H3K4me3/H3K4me1 как количественная мера предпочтения для данного локуса функционировать в качестве промотора или энхансера. Действительно, модификация лизина (K) в позиции 4 гистона 3 (H3), вызванная триметилированием или монометилированием маркирует, как было показано ранее, либо промоторные, либо энхансерные районы, соответственно. Таким образом, отношение геномных сигналов ChIP-seq этих маркеров (логарифм отношения высот пиков профилей) для конкретного геномного регуляторного участка показывает, промоторный или энхансерный это район. При этом не используется заранее заданная аннотация расположения промоторов, только экспериментальные геномные данные для исследуемых клеток.

Большинство не-контактирующих сайтов РНК-полимеразы II, проксимальных к старту транскрипции в базальной промоторной модели, показали высокий уровень логарифма отношения промоторного к энхансерному сигналов (Рисунок 5.21, панель 1; медиана равна 2.4; более 90% контактирующих участков имели значения логарифма отношения сигналов больше нуля) В то же время как большинство взаимодействующих сайтов РНК-полимеразы II, дистальных по отношению к старту транскрипции в одногенных моделях комплексов мультигенных комплексах (условно энхансерные

сайты), показали низкий уровень этого параметра (медиана < -0.72) логарифма отношения сигналов H3K4me3/H3K4me1 (Рисунок 5.21, панель 6). Таким образом, логарифм отношения сигналов H3K4me3 и H3K4me1 отражает относительные возможности промоторов и энхансеров.

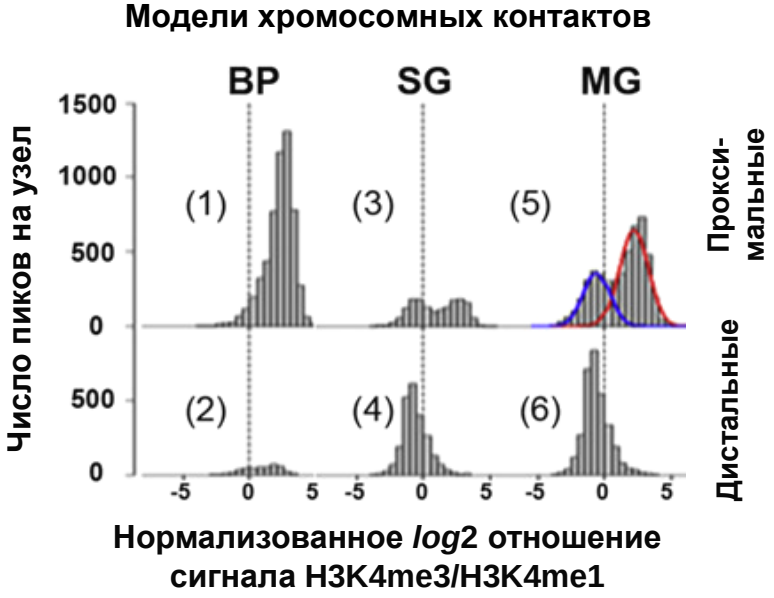


Рис. 5.21. Распределение моделей ChIA-PET контактов по отношению к промоторного к энхансерному сигналов в геноме человека [12].

Интересно отметить, что для участков взаимодействий РНК-полимеразы II, проксимальных к известным точкам старта транскрипции генов в мультигенных комплексах (Рисунок 5.21, панель 5), выявлено два пика в гистограмме логарифма отношения сигналов, что позволяет предположить присутствие как энхансерных, так и промоторных элементов, функционирующих в известных промоторных районах генов. То есть, некоторые промоторные участки не вносят вклад в транскрипцию с ближайшего старта транскрипции, но являются элементами регуляции других соседних генов. Таким образом, значительная доля контактирующих промоторов потенциально может иметь энхансерные функции.

Взаимодействующие промоторы могут обладать комбинаторными регуляторными функциями

Для исследования потенциальной энхансерной активности промоторов, была выполнена проверка с помощью люциферазной (luciferase) репортерной генной конструкции, используя вставку одного и нескольких промоторных элементов, по установленному методу характеризации промоторов и энхансеров [348]. В этом тестировании фрагменты размером около 500 нт из предполагаемых промоторных районов были клонированы перед репортерным геном люциферазы либо в проксимальной позиции как работающий промотор, либо в дистальной позиции как

предполагаемый энхансер, затем эти конструкты были внедрены (трансфектированы) в клетки линии MCF-7. Как показано на рисунке 5E, два взаимодействующих локуса *INTS1* и *MAFK* находятся на расстоянии 26Кб друг от друга на хромосоме 7, и данные RNA-Seq подтверждают активность этих двух генов в клетках линии MCF-7. Тем не менее, нормализованное отношение логарифма сигнала H3K4me3/H3K4me1 равно 0.36 для промотора гена *INTS1* и составляет 1.13 для промотора *MAFK*, позволяя предположить, что промотор *INTS1* может иметь энхансерные свойства. Для тестирования этого предположения, был клонирован фрагмент промотора *INTS1* в обоих ориентациях перед промотором *MAFK*, который был фланкирован геном люциферазы (luciferase - фермент монооксигеназа, катализирующая реакцию восстановленного люциферина с АТФ. Продукт реакции, аденилат, при окислении дает световой сигнал). Тест репортерного гена люциферазы дал как минимум семикратное увеличение 7 экспрессии люциферазы с промотора *MAFK* активированного фрагментом промотора *INTS1*, указывая что промоторы могут регулировать активность других промоторов, по меньшей мере, некоторые из промоторов (примеры приведены в статье [12], см. пример промотора *CALM1*).

Далее, исследовался вопрос, могут ли промоторы с энхансерной активностью действовать специфически. Были экспериментально поменяны местами промоторные элементы в двух примерах *INTS1*-и-*MAFK* и *C14orf102*-и-*CALM1* для дополнительного тестирования конструкций репортерных генов. Интересно отметить, что при размещении перед промотором *CALM1*, промотор *INTS1* показал значительное усиление промоторной активности *CALM1*. Подобным образом, комбинированная конструкция промотора *C14orf102* и энхансера *CALM1* также значительно увеличила промоторную активность *MAFK*. В то же время, промотор с удаленным TATA боксом и другие контрольные промоторы (как активные так и не активные), взятые от соседних генов, которые не вовлечены в промотор-промоторные взаимодействия, не показали кооперативного усиления к промоторной активности *MAFK* и *CALM1*. Таким образом, эти результаты позволяют предположить общность свойств для промоторов с энхансерной активностью, которая влияет на другие промоторы, вовлеченные во взаимодействия хроматина.

Дополнительно был клонирован фрагмент промотора с энхансерными свойствами (INTS1) в позицию, проксимальную к гену люциферазы, и вставлен сильный промотор (MAFK) с высоким уровнем лог-отношения K4me3/me1, характеризующим промоторные свойства, в дистальную позицию в репортерной конструкции гена. Промоторный фрагмент MAKF не показал значительной

энхансерной активности. Более того, низкое отношение H3K4me3/me1 («энхансерность») в энхансерном положении и высокое отношение H3K4m3/me1 («промоторность») в промоторном положении показали сигнал активации, но не наоборот [12].

Заметим, что необходимо дальнейшее исследование такой регуляторной взаимозаменяемости между промоторами. Недавняя статья [198] показала возможность компьютерного предсказания тканеспецифичных энхансеров по свойствам промоторов (комбинациям сайтов связывания), экспериментального исследования взаимозаменяемости не было представлено.

Клеточная специфичность дальнедействующих взаимодействий хроматина

Для анализа специфичности взаимодействий хроматина к типу клеток, контакты хроматина были определены максимально полно посредством глубокого секвенирования большего числа реплик ChIA-PET экспериментов для клеточных линий MCF-7 и K562. Пополненные библиотеки были хорошо воспроизводимы для детектируемых взаимодействий и таким образом достаточно надежны для сравнительного анализа между клеточными линиями. Эти библиотеки имели такую же структуру геномных характеристик, как и первые (пилотные) библиотеки. С помощью детальных наборов данных ChIA-PET и RNA-Seq был выполнен сравнительный анализ между двумя клеточными линиями и определена линие-специфичная генная экспрессия и контакты хроматина. Большинство генов, которые были специфично экспресссированы в соответствующих клетках, имели также и клеточно-специфичные контакты хроматина, подтверждая, что линие-специфичные взаимодействия хроматина обеспечивают структурную основу для клеточно-специфичной транскрипции. Анализ генных онтологий (ГО) выявил значимое обогащение эритроидных терминов ГО, таких как ответ на стимулы, циркуляция крови, иммунный ответ и развитие мезодермы для генов со специфической экспрессией и контактами хроматина в клетках линии K562, в то время как термины ГО, такие как развитие эктодермы и соответствующие биологические процессы, были отмечены в категориях онтологий для генов со специфичной экспрессией и контактами хроматина в клетках MCF-7. Как можно было ожидать, гены, общие для обеих клеточных линий, показали обогащение функциями домашнего хозяйства, такими как метаболизм, клеточный цикл и передача сигнала.

Среди взаимодействий хроматина, специфичных для клеток линии K562 были найдены многие ранее охарактеризованные взаимодействия, включая α- и β-глобиновые локусы [490, 508]. С помощью данных ChIA-PET были найдены

взаимодействия между локусом гена α-глобина и гиперчувствительными сайтами разрезания ДНКазы (DNase hyper-sensitive sites), представленными в области гена *C16orf35*, что согласуется с ранее опубликованными данными [508]. Кроме того, было установлено, что α-глобиновый локус расширяет взаимодействия на соседние домены, которые постоянно активны в обеих линиях клеток - K562 и MCF-7, тогда как контакты к α-глобиновым генам специфичны для K562, что позволяет предположить существование комплекса архитектуры хроматина для возможной пространственно-временной регуляции как конститутивной, так и линие-специфичной транскрипции. Подобным образом, локус β-глобиновых генов на хромосоме 11 также показал ранее известные K562-специфичные взаимодействия с окружающим локус-контролирующим районом (LCR - locus control region). Ген GREB1, специфичный для линии MCF-7, также был найден специфично ChIA-PET для MCF-7.

Ассоциированные с заболеваниями некодирующие элементы могут связываться с промоторами генов-мишеней через дальнедействующие взаимодействия

Данные показали, что энхансер-промоторные взаимодействия значимо обогащены по сравнению с другими типами взаимодействий для линие-специфичных генов в сравнении с генами, одинаково экспрессирующимися в обеих клеточных линиях. Это наблюдение поддерживает общую оценку того, что удаленно действующие энхансеры имеют тенденцию быть специфически вовлеченными в контакты хроматина к ткане-специфичным генам, что соответствует анализу по «широте» экспрессии в различных тканях для мультигенных и одногенных комплексов. Хотя потенциальные энхансеры могут быть определены экспериментально с помощью высокопроизводительных подходов [182], остается нерешенной проблема сопоставления энхансеров и их генов-мишеней, находящихся на удалении сотен тысяч нуклеотидов Более того, многие дальние энхансеры могут быть вложены в интронные районы других дистально расположенных генов [183], исключительно затрудняя соотнесение энхансеров их генам-мишеням. С помощью ChIA-PET было идентифицировано приблизительно 1000 ультра-дальнедействующих (расположенных далее 500Кб) энхансер-промоторных взаимодействий, которые также специфичны к исследованным клеточным линиям. Большая доля этих энхансеров взаимодействует с удаленными генами-мишенями через контакты с ближайшим промотором (промоторами).

Интересный пример представляет ген *SHH*, который экспрессируется в клетках MCF7, но не в клетках K562 (Рис. 5.22).

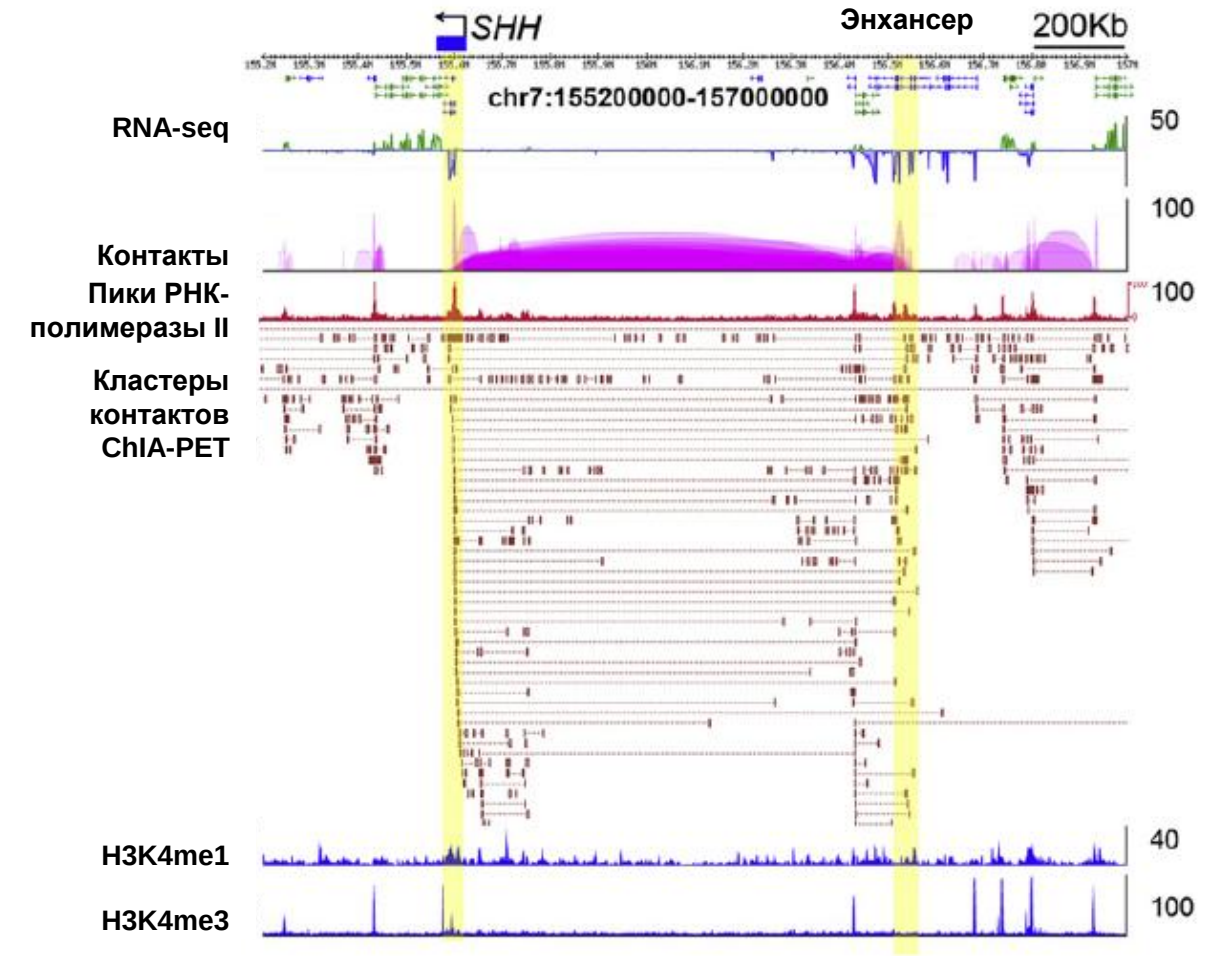


Рис. 5.22. Дальнодействующие взаимодействия между геном *SHH* (выделено желтым цветом, слева) и его энхансером расположенным в удалении 1 Мб в интроне LMBR1 (выделено желтым, справа). Экспрессия *SHH* наблюдается специфически в клетках линии MCF-7.

SHH важен для развития организма и соотносится с определенными видами рака [180]. Транскрипция *SHH* контролируется его энхансером, который расположен на расстоянии 1Мб и вложен в интронный район *LmbR1*. Точечная мутация в этом энхансере вызывает преаксиальную полидактилию (*preaxial polydactyly)* - общее врожденное нарушение формирования конечностей у млекопитающих [180]. С помощью ChIA-PET были найдены взаимодействия между промотором *SHH* и ранее охарактеризованным энхансерным сайтом *SHH* в интронном районе *LmbR1* в клетках линии MCF-7, но не было найдено взаимодействий в клетках K562, что соотносится с транскрипционным статусом *SHH*. Этот факт подтверждается более ранними наблюдениями [179].

В другом примере дальнодействующих взаимодействий выявленных с помощью ChIA-PET, было найдено два основных сайта хромосомных контактов расположенных на расстоянии приблизительно 600 Кб и около 1Мб о промотора гена *IRS1*, который участвует в развитии сахарного диабета второго типа. Этот ген специфически

экспрессировался в клетках MCF-7 в проведенном эксперименте. Недавние исследования полногеномных ассоциаций GWAS установило, что кластер SNP, локализованный в одном из энхансерных сайтов *IRS1*, генетически ассоциирован с риском устойчивости к инсулину, диабету второго типа и болезнью коронарной сердечной артерии [520].

Таким образом, ChIA-PET данные дают экспериментальное доказательство того, что этот локус риска развития заболевания может быть физически связан с промотором *IRS1*, потенциально работая как критичный дальнодействующий энхансер, регулирующий экспрессию *IRS1*, подобно описанному локусу *SHH*. Точечные мутации в этом районе у пациентов с сахарным диабетом второго типа могут нарушать энхансерную функцию и, следовательно, влиять на транскрипцию *IRS1*.

5.5. Заключение к Главе

С помощью полногеномного картирования PET кластеров были детально проанализированы ассоциированные с РНК-полимеразой II дальнодействующие взаимодействия хроматина.

Матерериалы данной Главы подтверждают следующее положение, выносимое на защиту:

Геномные области хромосомных контактов, опосредованных комплексом РНК-полимеразы II, обогащены сайтами связывания транскрипционных факторов и участками модификаций гистонов, связанными с активацией экспрессии генов.

Одна из интересных находок - обнаружение межгенных промотор-промоторных взаимодействий. Хотя кооперативные дальнедействующие взаимодействия среди разделенного NFkB были отмечены ранее [521], в эксперименте ChIA-PET впервые были найдены широко распространенные промотор-промоторные взаимодействия среди проксимальных и дистальных генов, указывающие на то, что в клетках этот механизм является общим. Экспериментальная проверка с помощью репортерных генных конструкций и нокдаун генов посредством малых интерферирующих РНК (siRNA) дает экспериментальное доказательство того, что некоторые промоторы в мультигенных комплексах могут кооперативно усиливать активность других промоторов, с которыми они контактируют. Эти наблюдения размывают условное определение промотора и регуляторных элементов транскрипции. Мультигенные комплексы, описанные в данном исследовании, в принципе подобны оперонам в

бактериях как механизм координированной транскрипционной активности родственных генов, предполагая возможность существования оперонных механизмов, связанных с хроматином (хро-оперон или хроперон (chro-operon или chroperon в английском варианте) для пространственно-временной регуляции транскрипции генов в ядрах клеток эукариот. Альтернативно, эти взаимодействия могут отражать стохастическое движение проксимальных и дистальных активных генов к локализованным транскрипционным фабрикам. Дальнейшие исследование пространственной архитектуры, найденные с помощью ChIA-PET, улучшат понимание транскрипционной регуляции в нормальных клетках человека и клетках в условиях развития заболеваний.

Основным выводом исследования является иерархичность организации топологических доменов в геноме человека, высокая степень близости регуляторных последовательностей к транскрибируемым генам в пространстве клеточного ядра. Впервые построены карты хромосомных контактов, опосредованных комплексом РНК-полимеразы II и рецептором эстрогена ERα в геноме человека, полученные по методу секвенирования ChIA-PET. Представлена классификация групп генов, находящихся в транскрипционных доменах, в зависимости от структуры контактов (хромосомных петель).

Результаты, представленные в настоящей главе, позволяют аргументировать следующий вывод:

Впервые выполнен компьютерный анализ карт хромосомных контактов, опосредованных рецептором эстрогенов ERα и комплексом РНК-полимеразы II в геноме человека, полученных с помощью технологии секвенирования ChIA-PET. Представлена классификация групп генов, находящихся в транскрипционных доменах, в зависимости от структуры контактов (хромосомных петель). Показано присутствие в участках хромосомных контактов, опосредованных комплексом РНК-полимеразы II, сайтов связывания различных транскрипционных факторов, определенных с помощью технологии ChIP-seq в геноме. Показана положительная корреляция участков хромосомных контактов с модификациями гистонов, характеризующими открытое состояние хроматина (H3K4me3, H3K9ac, H3K4me1).

ЗАКЛЮЧЕНИЕ И ОБСУЖДЕНИЕ

Программные средства, разработанные в рамках настоящей диссертационной работы, позволили получить новые теоретические результаты по анализу данных полногеномного секвенирования, контекстной структуры нуклеотидных последовательностей геномов, хромосомных контактов.

В соответствии с целью и задачами исследования разработаны компьютерные программы для определения позиций сайтов связывания транскрипционных факторов в геноме на основе технологий данных секвенирования, сопряженных с иммунопреципитацией хроматина (технологий ChIP-PET и ChIP-seq). Программы были написаны на языке C++, часть скриптов подготовлена в среде программирования R.

С помощью этих программ определены позиции сайтов в геноме, исследовано распределение сайтов относительно генов, выделены гены-мишени регуляторного воздействия транскрипционных факторов Oct4, Nanog, Sox2, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши и факторов ERα, MYC, PRDM14 в геноме человека [3, 13, 41, 42, 54].

Для оценки полноты экспериментов ChIP-seq создана компьютерная статистическая модель распределения сайтов связывания в геноме и написана программа, рассчитывающая по данным ChIP-seq число потенциальных сайтов в зависимости от достигнутой глубины секвенирования, процента специфичных прочтений и размера хромосом [38]. Впервые представлена аппроксимация числа сайтов и оценка сатурации (полноты) эксперимента ChIP-seq для транскрипционных факторов плюрипотентности в геноме мыши [3].

Для интеграции данных по регуляции и экспрессии генов разработан компьютерный метод определения генов-мишеней действия транскрипционных факторов по геномным координатам сайтов связывания и полногеномным данным изменения экспрессии генов. Метод применен для реконструкции регуляторных генных сетей транскрипции в эмбриональных стволовых клетках мыши по данным ChIP-seq [3]. Анализ экспрессии генов генома требует обработки данных микрочипов и решение задач оценки достоверности измерения сигнала. Выполнена оценка качества экспрессионных микрочипов (наборов проб) платформы Affymetrix U133 для генов человека; оценки качества наборов проб представлены в базе данных [46, 47].

Работа с полученными картами сайтов связывания различных транскрипционных факторов в эмбриональных стволовых клетках (ЭСК) мыши и человека позволила исследовать механизмы увеличения эффективности репрограммирования, найти новые закономерности организации сайтов в дистальных

энхансерах. Впервые показана ко-локализация сайтов, связывания факторов, относящихся к группе ключевых регуляторов плюрипотентности Oct4-Nanog-Sox2, для полногеномного распределения сайтов связывания транскрипционных факторов c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2 в ЭСК мыши [3, 39-41, 52, 54]. Впервые построены тепловые карты (термокарты) ко-локализации указанных сайтов в геноме мыши. Интересно отметить последовательное уточнение карт связывания и термокарт ко-локализации в том же типе клеток для различных факторов, опубликованное в независимых работах с участием автора, начиная с ТФ Oct4, Nanog, Sox2 [3], продолжая Eset [39], Nr5a2 [40], Tbx3 [41], Smad2 [54].

Впервые построены геномные карты сайтов связывания транскрипционного фактора Smad2 в эмбриональных стволовых клетках мыши [54]. Показано различие профилей связывания и набора генов-мишеней Smad2 в зависимости от дозы активатора. Впервые показана вариабельность геномного окружения выявляемых сайтов Smad2 в зависимости от дозы активатора, выделены нуклеотидные мотивы ко-факторов. Компьютерная обработка данных ChIP-seq в парных экспериментах была продолжена на модели рыбы *D.rerio* для ТФ Zic3 [43].

Для ЭСК человека данные ChIP-seq позволили установить сходные наборы совместно локализующихся сайтов факторов плюрипотентности [52]. Реконструирован список генов, ответственных за поддержание плюрипотентности в эмбриональных стволовых клетках человека в эксперименте с последовательным нокаутом транскриптов, и построена генная сеть их взаимодействия [42]. Впервые показана значимость транскрипционного фактора PRDM14 для поддержания плюрипотентного состояния клеток как регулятора OCT4; определен мотив связывания транскрипционного фактора PRDM14. Показана ассоциация сайтов связывания PRDM14 в геноме человека с кластерами сайтов транскрипционных факторов, ответственных за поддержание плюрипотентности OCT4-NANOG-SOX2, построена термокарта ко-локализации [42].

Построены геномные карты связывания факторов MYC, ERα, FOXA1 в геноме человека для различных клеточных линий опухолей [9, 13]. Установлено соответствие аффинности связывания транскрипционного фактора ERα, и числа прочтений ДНК в экспериментах иммунопреципитации хроматина для ERα (ChIP-seq и ChIP-PET) в культуре клеток MCF-7 [13].

Показано, что присутствие маркеров открытого хроматина (отсутствия нуклеосомной упаковки FAIRE) и маркеров модификаций хроматина, в частности гистона H3 (H3K4me3, H3K4me1, H3K9ac, H3K14ac), определенных с помощью

технологии ChIP-seq, позволяет предсказать с высокой точностью сайты связывания транскрипционного фактора ERα в геноме человека [13]. Оптимизирована компьютерная модель предсказания сайтов связывания ERα в геноме на основе данных о маркерах хроматина, полученных с помощью ChIP-seq.

Анализ структуры хроматина опирается на определение позиций нуклеосом в геноме. Показана связь положения нуклеосом, позиции которых в геноме определены с помощью полногеномного секвенирования, и связывания транскрипционных факторов в промоторных районах генов, установленного с помощью технологии ChIP-chip на примере генома дрожжей [51, 62]. Исследование предпочтений в позиционировании нуклеосомных сайтов продолжает работы по предсказанию положения нуклеосом только на основании контекстов нуклеотидных последовательностей [50].

Исследование сайтов связывания онкогена ERα в геноме человека продолжалось с использованием более совершенных технологий, позволяющих определять не только участки связывания на хромосомах, но и пары таких участков, пространственно сближенных в ядре клетке, устанавливая тем самым трехмерную структуру хромосомных контактов. Впервые построены карты хромосомных контактов, опосредованных комплексом РНК-полимеразы II и транскрипционным фактором ERα в геноме человека, полученные по методу секвенирования ChIA-PET [12, 21]. Представлена классификация групп генов, находящихся в транскрипционных доменах, в зависимости от структуры контактов (хромосомных петель).

Впервые показана ассоциация участков хромосомных контактов, опосредованных комплексом РНК-полимеразы II, с сайтами связывания транскрипционных факторов, в том числе в дистальных регуляторных районах, и с изучавшимися ранее модификациями гистонов, характеризующими открытое состояние хроматина (H3K4me3, H3K9ac, H3K4me1), определенными с помощью технологии ChIP-seq в масштабе генома [12].

Исследование хромосомных контактов не только подтверждает иерархичность организации топологических доменов в геноме человека и близость регуляторных последовательностей к транскрибируемым генам в пространстве ядра, но и позволяет по-новому изучать регуляцию экспрессии генов, через корреляции консервативных элементов, общие категории генных онтологий, паттерны совместной экспрессии.

ВЫВОДЫ ПО ДИССЕРТАЦИОННОЙ РАБОТЕ

1) Впервые разработан подход для статистической оценки нижней и верхней границ общего числа сайтов связывания транскрипционных факторов в геноме мыши на основе анализа экспериментальных данных ChIP-seq. Этот подход дает возможность оценки качества экспериментов ChIP-seq для выявления сайтов связывания транскрипционных факторов при заданном объеме секвенирования и размере генома.

2) Разработаны компьютерные методы и программы для анализа данных по полногеномному секвенированию, сопряженному с иммунопреципитацией хроматина, получаемых в экспериментах ChIP-PET и ChIP-seq, и распознавания на этой основе сайтов связывания транскрипционных факторов в геномах человека, мыши, рыбы *Danio rerio*.

3) С помощью компьютерного анализа данных экспериментов ChIP-seq на эмбриональных стволовых клетках мыши впервые построена термокарта совместной локализации транскрипционных факторов Oct4, Nanog, Sox2, Klf4, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши. Показана совместная геномная локализация сайтов связывания транскрипционных факторов Oct4, Nanog, Sox2 и Klf4, относящихся к ключевым регуляторам плюрипотентности.

4) Впервые по данным экспериментов ChIP-seq на эмбриональных стволовых клетках мыши определены группы сайтов связывания транскрипционного фактора Smad2 в условиях активации и подавления экспрессии гена Smad2 под действием внешних факторов - белка Activin и ингибитора SB431542, соответственно. В геномном окружении сайтов Smad2 найдены специфичные группы нуклеотидных мотивов, соответствующих потенциальным сайтам связывания других транскрипционных факторов.

5) На основе компьютерной модели эксперимента с последовательным подавлением транскрипции генов в эмбриональных стволовых клетках (ЭСК) человека показана роль транскрипционного фактора PRDM14 в поддержании плюрипотентности. Для транскрипционного фактора PRDM14 по данным ChIP-seq найдены его гены-мишени в ЭСК человека, включающие OCT4. Впервые определена структура сайта связывания PRDM14.

6) С помощью компьютерного анализа данных экспериментов ChIP-seq в ЭСК человека построена термокарта расположения кластеров сайтов связывания для

транскрипционных факторов OCT4, NANOG, SOX2 и PRDM14 в геноме человека. Показано совместное геномное расположение сайтов связывания транскрипционных факторов OCT4, NANOG, SOX2 в ЭСК человека, аналогичное расположению сайтов связывания их гомологов в ЭСК мыши.

7) Установлена положительная взаимосвязь (p<0.001) между силой связывания транскрипционных факторов ERα и MYC, измеренной с помощью количественной ПЦР, и числом прочтений ДНК в экспериментах иммунопреципитации хроматина ChIP-PET и ChIP-seq для транскрипционных факторов MYC и ERα, в культурах клеток опухолей человека P493 и MCF-7, соответственно. Выявлены нуклеотидные мотивы транскрипционных факторов, связывающихся в окрестностях сайтов ERα.

8) Рассчитаны позиции положения нуклеосом в геноме дрожжей на основе данных секвенирования защищенных нуклеосомой фрагментов ДНК. Показано, что сайты связывания транскрипционных факторов в промоторных районах генов дрожжей, определенные с помощью технологии ChIP-chip, свободны от нуклеосомной упаковки.

9) Показано, что присутствие маркеров открытого хроматина и маркеров модификаций гистонов, в частности гистона H3 (H3K4me3, H3K4me1, H3K9ac, H3K14ac), определенных с помощью технологии ChIP-seq, позволяет предсказать с высокой точностью сайты связывания транскрипционного фактора ERα в геноме человека.

10) Впервые выполнен компьютерный анализ карт хромосомных контактов, опосредованных рецептором эстрогенов ERα и комплексом РНК-полимеразы II в геноме человека, полученных с помощью технологии секвенирования ChIA-PET. Представлена классификация групп генов, находящихся в транскрипционных доменах, в зависимости от структуры контактов (хромосомных петель). Показано присутствие в участках хромосомных контактов, опосредованных комплексом РНК-полимеразы II, сайтов связывания различных транскрипционных факторов, определенных с помощью технологии ChIP-seq в геноме. Показана положительная корреляция участков хромосомных контактов с модификациями гистонов, характеризующими открытое состояние хроматина (H3K4me3, H3K9ac, H3K4me1).

Список публикаций по теме диссертации

Статьи в научных журналах

1. **Орлов Ю.Л.**, Левицкий В.Г., Смирнова О.Г., Подколодная О.А., Хлебодарова Т.М., Колчанов Н.А. **(2006)** Статистический анализ последовательностей ДНК, содержащих сайты формирования нуклеосом. // *Биофизика* 51:608-14.
2. Воробьева Н.В., Билтуева Л.С., **Орлов Ю.Л.**, Графодатский А.С., Колчанов Н.А. **(2006)** Интерстициальные теломерные повторы, как маркеры эволюционных преобразований кариотипа млекопитающих: хромосома 2 человека. // *Биофизика* 51:602-7.
3. **Orlov Y.L.**, Te Boekhorst R., Abnizova I.I. **(2006)** Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. // *J Bioinform Comput Biol.* 4:523-36.
4. Zeller K.I., Zhao X., Lee C.W., Chiu K.P., Yao F., Yustein J.T., Ooi H.S., **Orlov Y.L.**, Shahab A., Yong H.C., Fu Y., Weng Z., Kuznetsov V.A., Sung W.K., Ruan Y., Dang C.V., Wei C.L. **(2006)** Global mapping of c-Myc binding sites and target gene networks in human B cells. // *Proc Natl Acad Sci U S A.* 103:17834-9.
5. Zhao X.D., Han X., Chew J.L., Liu J., Chiu K.P., Choo A., **Orlov Y.L.**, Sung W.K., Shahab A., Kuznetsov V.A., Bourque G., Oh S., Ruan Y., Ng H.H., Wei C.L. **(2007)** Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. // *Cell Stem Cell*. 1(3):286-98.
6. **Orlov Y.L.**, Zhou J., Lipovich L., Shahab A., Kuznetsov V.A. **(2007)** Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis. *In Silico Biol*. 7(3):241-60.
7. Kuznetsov V.A., **Orlov Y.L.**, Wei C.L., Ruan Y. **(2007)** Computational analysis and modeling of genome-scale avidity distribution of transcription factor binding sites in chip-pet experiments. *Genome Inform.* 19:83-94.
8. Chen X., Xu H., Yuan P., Fang F., Huss M., Vega V.B., Wong E., **Orlov Y.L.**, Zhang W., Jiang J., Loh Y.H., Yeo H.C., Yeo Z.X., Narang V., Govindarajan K.R., Leong B., Shahab A., Ruan Y., Bourque G., Sung W.K., Clarke N.D., Wei C.L., Ng H.H. **(2008)** Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. // *Cell*. 133(6):1106-17.

9. Fullwood M.J., Liu M.H., Pan Y.F., Liu J., Xu H., Mohamed Y.B., **Orlov Y.L.**, Velkov S., Ho A., Mei P.H., Chew E.G., Huang P.Y., Welboren W.J., Han Y., Ooi H.S., Ariyaratne P.N., Vega V.B., Luo Y., Tan P.Y., Choy P.Y., Wansa K.D., Zhao B., Lim K.S., Leow S.C., Yow J.S., Joseph R., Li H., Desai K.V., Thomsen J.S., Lee Y.K., Karuturi R.K., Herve T., Bourque G., Stunnenberg H.G., Ruan X., Cacheux-Rataboul V., Sung W.K., Liu E.T., Wei C.L., Cheung E., Ruan Y. **(2009)** An oestrogen-receptor-alpha-bound human chromatin interactome. // *Nature*. 462(7269):58-64.
10. Yuan P., Han J., Guo G., **Orlov Y.L.**, Huss M., Loh Y.H., Yaw L.P., Robson P., Lim B., Ng H.H. **(2009)** Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. // *Genes Dev.* 23(21):2507-20.
11. Grinchuk O.V., Jenjaroenpun P., **Orlov Y.L.**, Zhou J., Kuznetsov V.A. **(2010)** Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns. // *Nucleic Acids Res.* 38(2):534-47.
12. Heng J.C., Feng B., Han J., Jiang J., Kraus P., Ng J.H., **Orlov Y.L.**, Huss M., Yang L., Lufkin T., Lim B., Ng H.H. **(2010)** The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. // *Cell Stem Cell*. 6(2):167-74.
13. Han J., Yuan P., Yang H., Zhang J., Soh B.S., Li P., Lim S.L., Cao S., Tay J., **Orlov Y.L.**, Lufkin T., Ng H.H., Tam W.L., Lim B. **(2010)** Tbx3 improves the germ-line competency of induced pluripotent stem cells. // *Nature*. 463(7284):1096-100.
14. Goh W.S., **Orlov Y.**, Li J., Clarke N.D. **(2010)** Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. // *PLoS Comput Biol.* 6(1):e1000649.
15. Guo X., Popadin K.Y., Markuzon N., **Orlov Y.L.**, Kraytsberg Y., Krishnan K.J., Zsurka G., Turnbull D.M., Kunz W.S., Khrapko K. Repeats, longevity and the sources of mtDNA deletions: evidence from 'deletional spectra' // *Trends Genet.* 2010 26(8):340-343.
16. Chia N.-Y., Chan Y.-S., Feng B., Lu X., **Orlov Y.L.**, Moreau D., Kumar P., Yang L., Jiang J., Lau M.-S., Huss M., Soh B.-S., Kraus B.-S., Lufkin T., Lim B., Clarke N., Bard F., Ng H.H. **(2010)** A genome-wide RNAi screen identifies PRDM14 as a regulator of POU5F1 and human embryonic stem cell identity // *Nature*. 468(7321): 316-20.
17. Heng J.C., **Orlov Y.L.**, Ng H.H. **(2010)** Transcription Factors for the Modulation of Pluripotency and Reprogramming. // In: *Cold Spring Harb Symp Quant Biol. 2010*; 75:237-44.
18. Joseph R., **Orlov Y.L.**, Huss M., Sun W., Kong S.L., Ukil L., Pan Y.F., Li G., Lim M., Thomsen J.S., Ruan Y., Clarke N.D., Prabhakar S., Cheung E., Liu E.T. **(2010)** Integrative model of genomic factors for determining binding site selection by estrogen receptor α. // *Mol Syst Biol*. 6:456.

19. **Орлов Ю.Л.**, Ефимов В.М., Орлова Н.Г. **(2011)** Статистические оценки экспрессии мобильных элементов в геноме человека на основе клинических данных экспрессионных микрочипов. // *Вавиловский журнал генетики и селекции.* 15(2): 327-339.
20. Lee K.L., Lim S.K., **Orlov Y.L.**, Yit le Y., Yang H., Ang L.T., Poellinger L., Lim B. **(2011)** Graded Nodal/Activin signaling titrates conversion of quantitative phospho-Smad2 levels into qualitative embryonic stem cell fate decisions. // *PLoS Genet*. 7(6):e1002130.
21. Putta P., **Orlov Y.L.**, Podkolodnyy N.L., Mitra C.K. (2011) Relatively conserved common short sequences in transcription factor binding sites and miRNA. // *Вавиловский журнал генетики и селекции*. Том 15, № 4, Стр. 750-756.
22. Li G., Ruan X., Auerbach R.K., Sandhu K.S., Zheng M., Wang P., Poh H.M., Goh Y., Lim J., Zhang J., Sim H.S., Peh S.Q., Mulawadi F.H., Ong C.T., **Orlov Y.L.**, Hong S., Zhang Z., Landt S., Raha D., Euskirchen G., Wei C.L., Ge W., Wang H., Davis C., Fisher-Aylor K.I., Mortazavi A., Gerstein M., Gingeras T., Wold B., Sun Y., Fullwood M.J., Cheung E., Liu E., Sung W.K., Snyder M., Ruan Y. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. // *Cell*. 148(1-2):84-98.
23. **Orlov Y.**, Xu H., Afonnikov D., Lim B., Heng J.C., Yuan P., Chen M., Yan J., Clarke N., Orlova N., Huss M., Gunbin K., Podkolodnyy N., Ng H.H. (2012) Computer and Statistical Analysis of Transcription Factor Binding and Chromatin Modifications by ChIP-seq data in Embryonic Stem Cell // *J Integr Bioinform*. 9(2):211.
24. Кожевникова О.С., Мартыщенко М.К., Генаев М.К., Корболина М.К., Муралева Н.А., Колосова Н.А., **Орлов Ю.Л.** (2012) RatDNA: база данных микрочиповых исследований на крысах для генов, ассоциированных с заболеваниями старения // *Вавиловский журнал генетики и селекции*. 16(4/1): 756-765.
25. **Орлов Ю.Л.**, Брагин А.О., Медведева И.В., Гунбин И.В., Деменков П.С., Вишневский О.В., Левицкий В.Г., Ощепков В.Г., Подколодный Н.Л., Афонников Д.А., Гроссе И., Колчанов Н.А. (2012) ICGenomics: программный комплекс анализа символьных последовательностей геномики // *Вавиловский журнал генетики и селекции*. 16(4/1): 732-741.
26. Баттулин Н.Р., Фишман В.С., **Орлов Ю.Л.**, Мензоров А.Г., Афонников Д.А., Серов О.Л. (2012) 3С-методы в исследованиях пространственной организации генома. // *Вавиловский журнал генетики и селекции*. 16(4/2): 872-878.
27. Winata C.L., Kondrychyn I., Kumar V., Srinivasan K.G., **Orlov Y.**, Ravishankar A., Prabhakar S., Stanton L.W., Korzh V., Mathavan S. (2013) Genome-wide analysis reveals Zic3 interaction with distal regulatory elements of stage specific developmental genes in zebrafish. // *PLOS Genetics* 9(10): e1003852.

28. Kozhevnikova O.S., Korbolina E.E., Stefanova N.A., Muraleva N.A., **Orlov Y.L.**, Kolosova N.G. **(2013)** Association of AMD-like retinopathy development with an Alzheimer's disease metabolic pathway in OXYS rats // *Biogerontology*. 14(6): 753-62.
29. Медведева И.В., Вишневский О.В., Сафронова Н.С., Кожевникова О.С., Генаев М.А., Кочетов А.В., Афонников Д.А., **Орлов Ю.Л.** (2013) Компьютерный анализ данных экспрессии генов в клетках мозга, полученных с помощью микрочипов и высокопроизводительного секвенирования // *Вавиловский журнал генетики и селекции*. Т. 17, № 4/1, С. 629-638.
30. **Орлов Ю.Л.** (2014) Компьютерное исследование регуляции транскрипции генов эукариот с помощью данных экспериментов секвенирования и иммунопреципитации хроматина // *Вавиловский журнал генетики и селекции.* Т. 18, № 1, С. 193-206.

Статьи в сборниках научных трудов

31. **Orlov Y.L.**, Zhou J.T., Chen J., Shahab A., Kuznetsov V.A. (2007) APMA Database for Affymetrix target sequences mapping, quality assessment and expression data mining. // In: *Pattern Recognition in Bioinformatics: second IAPR international workshop, PRIB 2007* (J.C. Ragapakse, B. Schmidt and G. Volkert, Eds), LNBI 4774 Springer-Verlag: Berlin-Heidelberg 2007; 166-177.
32. **Orlov Y.L.**, Huss M.E., Joseph R., Xu H., Vega V.B., Lee Y.K., Goh W.S., Thomsen J.S., Cheung E.C., Clarke N.D., Ng H.H. **(2009)** Genome-wide statistical analysis of multiple transcription factor binding sites obtained by ChIP-seq technologies. // In: *Proceedings of the 1st ACM Workshop on Breaking Frontiers of Computational Biology (CompBio '09).* 2009. ACM, New York, NY, 11-18.
33. **Орлов Ю.Л.**, Вишневский О.В., Витяев Е.Е., Кожевникова О.С., Афонников Д.А., Колчанов Н.А. (2013) Биоинформационный анализ экспрессии генов в клетках мозга. // XV Всероссийская научно-техническая конференция «Нейроинформатика-2013»: Сборник научных трудов. М.: НИЯУ МИФИ. 2013. С. 74-85.
34. Kolchanov N.A., **Orlov Y.L.** (2013) Introductory note for BGRS-2012 special issue. // *Journal of Bioinformatics and Computational Biology* Vol. 11, No. 1: 1302001.
35. Медведева И.В., Вишневский О.В., Сафронова Н.С., Кожевникова О.С., Суслов В.В., Кулакова Е.В., Спицына А.М., Афонников Д.А., Кочетов А.В., **Орлов Ю.Л.** (2014) Геномная организация и контекстные характеристики генов с повышенной экспрессией в клетках мозга // XVI Всероссийская научно-техническая конференция «Нейроинформатика-2014»: Сборник научных трудов. М.: НИЯУ МИФИ. Ч. 2., С. 32-42.

Глава в монографии

36. Подколодный Н.Л., Афонников Д.А., Гунбин К.В., Генаев М.А., **Орлов Ю.Л.**, Игнатьева Е.В., Вишневский О.В., Иванисенко В.А., Деменков П.С., Колчанов Н.А. Разработка методов, алгоритмов и программ параллельного моделирования в биоинформатике // В: Вычислительные методы, алгоритмы и аппаратурно-программный инструментарий параллельного моделирования природных процессов / М.Г. Курносов [и др.]; отв. ред. В.Г.Хорошевский; Рос. Акад. Наук, Сиб. Отд-ние, Ин-т физики полупроводников им. А.В. Ржанова [и др.]. – Новосибирск: Изд-во СО РАН, 2012. – 335с. – (Интеграционные проекты СО РАН; вып. 33) с. 271-334.

Тезисы конференций

37. Kuznetsov V.A., Zhou J.T., George J., **Orlov Yu.L. (2006)** Genome-wide co-expression patterns of human cis-antisense gene pairs // Proceedings of BGRS'2006, Novosibirsk, Inst. of Cytology&Genetics Press, V.1, p.90-93.
38. **Orlov Yu.L.**, Zhou J.T., Lipovich L., Yong H.C., Li Yi, Shahab A., Kuznetsov V.A. **(2006)** A comprehensive quality assessment of the affymetrix u133a&b probesets by an integrative genomic and clinical data analysis approach // Proceedings of BGRS'2006, Novosibirsk, Inst. of Cytology&Genetics Press, V.1, p.126-129.
39. **Orlov Y.L.**, Huss M., Vega V.B., Clarke N. **(2008)** Statistical issues in genome-wide transcription factor binding sites analysis based on chromatin IP (ChIP-seq) // In: Proceedings of BGRS'2008, Novosibirsk, Inst. of Cytology&Genetics Press, V.1, 179.
40. **Orlov Y.L.**, Chen D., Dobrovolskaya O., Meng Y., Chen L., Afonnikov D.A., Chen M. (2012) Computer analysis and database presentation of antisense transcription associated with microRNA targets in plant genomes // In: The Eighth International Conference on Bioinformatics of Genome Regulation and Structure \ System biology (BGRS\SB'2012) 25-29 июня 2012, издательство ИЦиГ СО РАН, г.Новосибирск. P.227.
41. **Orlov Y.L.**, Li G., Auerbach R., Sandhu K.S., Ruan X., Fullwood M.J., Podkolodnyy N.L., Afonnikov D.A., Liu E., Wei C.L, Snyder M., Ruan Y. (2012) 3D chromosome contacts and chromatin interactions revealed by sequencing // In: The Eighth International Conference on Bioinformatics of Genome Regulation and Structure \ System biology (BGRS\SB'2012) 25-29 июня 2012, издательство ИЦиГ СО РАН, г.Новосибирск. P.228.

42. **Orlov Y.L.**, Martyschenko M.K., Afonnikov D.A., Rasskazov D.A., Fomin E.S., Kuchin N.V., Glinsky B.M., Podkolodnyy N.L., Kolchanov N.A. (2012) Supercomputer applications in bioinformatics: Shared Facility Center “Bioinformatics” of Siberian Branch of the Russian Academy of Sciences // In: The Eighth International Conference on Bioinformatics of Genome Regulation and Structure \ System biology (BGRS\SB'2012) 25-29 июня 2012, издательство ИЦиГ СО РАН, г.Новосибирск. P.229.
43. **Orlov Y.L.**, Li G., Afonnokov D.A., Lim B., Clarke N., Huss M., Gunbin K.V., Ruan Y., Podkolodnyy N.L., Chen M., Ng H.-H. (2012) Transcription factor binding and chromatin modifications analysis by ChIP sequencing data // In: The Eighth International Conference on Bioinformatics of Genome Regulation and Structure \ System biology (BGRS\SB'2012) 25-29 июня 2012, издательство ИЦиГ СО РАН, г.Новосибирск. P.230.
44. Podkolodnyy N.L., Demenkov P.S., Gunbin .V., **Orlov Y.L.**, Fomin E.S., Alemasov N.A., Kazantsev F.V., Vishnevsky O.V., Ivanisenko V.A., Afonnikov D.A., Kuchin N.V., Glinsky B.M., Kolchanov N.A. (2012) High performance computing in bioinformatics: case studies // In: The Eighth International Conference on Bioinformatics of Genome Regulation and Structure \ System biology (BGRS\SB'2012) 25-29 июня 2012, издательство ИЦиГ СО РАН, г.Новосибирск. P. 243.
45. Safronova N.S., Suslov V.V., Afonnikov D.A., Podkolodnyy N.L., Mitra C.K., **Orlov Y.L.** (2012) Oligonucleotide frequencies and GC content of bacterial genomes are related to the environment evolution // In: The Eighth International Conference on Bioinformatics of Genome Regulation and Structure \ System biology (BGRS\SB'2012) 25-29 июня 2012, издательство ИЦиГ СО РАН, г.Новосибирск. P.276.
46. Matushkin Y.G., Levitsky V.G., **Orlov Y.L.**, Likhoshvai V.A. (2012) Correlation between transcription efficiency initiation and translation efficiency for Sacharromyces cereviciae and Schizosaccharomyces_pombe // In: The Eighth International Conference on Bioinformatics of Genome Regulation and Structure \ System biology (BGRS\SB'2012) 25-29 июня 2012, издательство ИЦиГ СО РАН, г. Новосибирск. P.200.
47. **Orlov Y.L.**, Dobrovolskaya O., Yuan C.H., Afonnikov D.A., Zhu Y., Chen M. (2012) Integrative computer analysis of antisense transcripts and miRNA targets in plant genomes // *Journal of Stress Physiology & Biochemistry* (ISSN 1997-0838), Vol. 8 No. 3, S.7.

48. **Orlov Y.**, Xu H., Afonnikov D., Lim B., Heng J.-C., Yuan P., Chen M., Yan J., Clarke N., Orlova N., Huss M., Gunbin K., Podkolodnyy N., Ng H.H. (2012) Computer and Statistical Analysis of Transcription Factor Binding and Chromatin Modifications by ChIP-seq data in Embryonic Stem Cell // In: Proceedings IB 2012 (International Symposium on Integrative Bioinformatics. The 8th Annual Meeting) Zhejiang University Press, Hangzhou, China, 80-92.
49. **Orlov Y.L.**, Li G., Sandhu K.S., Fullwood M.J., Afonnikov D.A., Wei C.L., Serov O.L., Kolchanov N.A., Ruan Y. (2012) Computer analysis of 3D chromosome contacts mediated by RNA Pol II in human // Международный симпозиум «SysPatho-Системная биология и медицина». 11-14 сентября 2012 г. в Санкт-Петербурге (Царское Село), С.31-32.
50. **Орлов Ю.Л.**, Ершов Н.И., Вината С.Л., Кондрихин И., Матаван И., Афонников Д.А., Колчанов Н.А. (2012) Полногеномный анализ генов развития, регулируемых транскрипционными факторами у мыши и *D.Rerio* на основе технологии ChIP-seq // Сборник тезисов: III международная научно-практическая конференция «Постгеномные методы анализа в биологии, лабораторной и клинической медицине» (Postgenome-2012), 22-24 ноября, г.Казань. С.126.
51. Safronova N.S., **Orlov Y.L.** (2013) Computer programs for complexity estimation and oligonucleotided analysis of regulatory DNA // In: Program and abstracts of International Young Scientist School “System biology and bioinformatics” (Novosibirsk 2013). Издание института цитологии и генетики СО РАН, Новосибирск 2013, с.39-40.
52. **Orlov Y.**, Afonnikov D., Battulin N., Serov O., Kolchanov N., Li G., Ruan Y. (2013) 3D organization of chromosomes mediated by RNAPII complex contacts in human genome. // In: Program of International Moscow Conference on Computational Molecular Biology MCCMB’2013, Moscow, Russia, ITTP RAS, Bioinformatic Seminar, p.14.

Всего по теме диссертации: 52 публикации, из них 30 статей (ВАК), 5 статей в сборниках научных трудов, одна глава в монографии, 16 тезисов конференций.

Список литературы

1. Venter J.C., Adams M.D., Myers E.W., et al. The sequence of the human genome. *Science* - 2001. - V. 291. - 5507. - p. 1304-51.
2. Liu L., Li Y., Li S., et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol* - 2012. - V. 2012. - p. 251364.
3. 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. - 6. - p. 1106-17.
4. Mortazavi A., Williams B.A., McCue K., et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* - 2008. - V. 5. - 7. - p. 621-8.
5. Tucker T., Marra M. and Friedman J.M. Massively parallel sequencing: the next big thing in genetic medicine. *Am J Hum Genet* - 2009. - V. 85. - 2. - p. 142-54.
6. Ewing A.D. and Kazazian H.H., Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* - 2010. - V. 20. - 9. - p. 1262-70.
7. Kedes L. and Campany G. The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE competition. *Nat Genet* - 2011. - V. 43. - 11. - p. 1055-8.
8. Bernstein B.E., Birney E., Dunham I., et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* - 2012. - V. 489. - 7414. - p. 57-74.
9. 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 U S A* - 2006. - V. 103. - 47. - p. 17834-9.
10. Collas P. and Dahl J.A. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Front Biosci* - 2008. - V. 13. - p. 929-43.
11. Malone B.M., Tan F., Bridges S.M., et al. Comparison of four ChIP-Seq analytical algorithms using rice endosperm H3K27 trimethylation profiling data. *PLoS One* - 2011. - V. 6. - 9. - p. e25260.
12. 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. - 1-2. - p. 84-98.
13. Joseph R., Orlov Y.L., Huss M., et al. Integrative model of genomic factors for determining binding site selection by estrogen receptor-alpha. *Mol Syst Biol* - 2010. - V. 6. - p. 456.
14. Malone J.H. and Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol* - 2011. - V. 9. - p. 34.
15. 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. - 1. - p. 207-19.
16. Kuznetsov V.A., Orlov Y.L., Wei C.L., et al. 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.
17. 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.
18. Park P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* - 2009. - V. 10. - 10. - p. 669-80.
19. Zhao X.D., Han X., Chew J.L., et al. Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* - 2007. - V. 1. - 3. - p. 286-98.

20. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* - 2007. - V. 8. - 4. - p. 286-98.
21. Fullwood M.J., Liu M.H., Pan Y.F., et al. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* - 2009. - V. 462. - 7269. - p. 58-64.
22. 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. - 1. - p. 90-8.
23. Marti-Renom M.A. and Mirny L.A. Bridging the resolution gap in structural modeling of 3D genome organization. *PLoS Comput Biol* - 2011. - V. 7. - 7. - p. e1002125.
24. Belton J.M., McCord R.P., Gibcus J.H., et al. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* - 2012. - V. 58. - 3. - p. 268-76.
25. Колчанов Н.А., *Теоретическое исследование закономерностей структурно-функциональной организации и эволюции генетических макромолекул*. 1988, Институт Цитологии и Генетики СО АН: Новосибирск. p. 542.
26. Pevzner P.A., *Computational Molecular Biology: An Algorithmic Approach*. 2000: MIT Press. 314.
27. Mount D.W., *Bioinformatics. Sequence and genome analysis*. 2001, New York: CSHL Press. 564.
28. Berger B., Peng J. and Singh M. Computational solutions for omics data. *Nat Rev Genet* - 2013. - V. 14. - 5. - p. 333-46.
29. Durbin R. E.S., Krogh A., Mitchson G., *Biological sequence analysis*. 1998, Cambridge: Cambridge University Press. 356.
30. Gusfield D., *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. 1997, Cambridge Cambridge University Press,. 530.
31. Франк-Каменецкий М.Д., ed. *Компьютерный анализ генетических текстов*. 1990, Наука: Москва. 267.
32. Kapushesky M., Emam I., Holloway E., et al. Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res* - 2010. - V. 38. - Database issue. - p. D690-8.
33. Wu C., Orozco C., Boyer J., et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* - 2009. - V. 10. - 11. - p. R130.
34. Barrett T., Troup D.B., Wilhite S.E., et al. NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res* - 2011. - V. 39. - Database issue. - p. D1005-10.
35. Flicek P., Amode M.R., Barrell D., et al. Ensembl 2014. *Nucleic Acids Res* - 2014. - V. 42. - 1. - p. D749-55.
36. Kuhn R.M., Haussler D. and Kent W.J. The UCSC genome browser and associated tools. *Brief Bioinform* - 2013. - V. 14. - 2. - p. 144-61.
37. Orlov Y., Xu H., Afonnikov D., et al. Computer and statistical analysis of transcription factor binding and chromatin modifications by ChIP-seq data in embryonic stem cell. *J Integr Bioinform* - 2012. - V. 9. - 2. - p. 211.
38. Orlov Y.L. H.M.E., Joseph R., Xu H., Vega V.B., Lee Y.K., Goh W.S., Thomsen J.S., Cheung E.C., Clarke N.D., Ng H.H. *Genome-wide statistical analysis of multiple transcription factor binding sites obtained by ChIP-seq technologies*. in *1st ACM Workshop on Breaking Frontiers of Computational Biology (CompBio '09)*. 2009. Italy: ACM, New York.
39. Yuan P., Han J., Guo G., et al. Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. *Genes Dev* - 2009. - V. 23. - 21. - p. 2507-20.

40. Heng J.C., Feng B., Han J., et al. The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. *Cell Stem Cell* - 2010. - V. 6. - 2. - p. 167-74.
41. Han J., Yuan P., Yang H., et al. Tbx3 improves the germ-line competency of induced pluripotent stem cells. *Nature* - 2010. - V. 463. - 7284. - p. 1096-100.
42. Chia N.Y., Chan Y.S., Feng B., et al. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature* - 2010. - V. 468. - 7321. - p. 316-20.
43. Winata C.L., Kondrychyn I., Kumar V., et al. Genome wide analysis reveals Zic3 interaction with distal regulatory elements of stage specific developmental genes in zebrafish. *PLoS Genet* - 2013. - V. 9. - 10. - p. e1003852.
44. Орлов Ю.Л. Брагин А.О., Медведева И.В., Гунбин И.В., Деменков П.С., Вишневский О.В., Левицкий В.Г., Ощепков В.Г., Подколодный В.Г., Афонников В.Г., Гроссе И., Колчанов Н.А. ICGenomics: программный комплекс анализа символьных последовательностей геномики *Вавиловский журнал генетики и селекции* - 2012. - V. 16. - 4/1. - p. 732-741.
45. Kozhevnikova O.S., Korbolina E.E., Stefanova N.A., et al. Association of AMD-like retinopathy development with an Alzheimer's disease metabolic pathway in OXYS rats. *Biogerontology* - 2013. - V. -
46. Orlov Y.L., Zhou J., Lipovich L., et al. Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis. *In Silico Biol* - 2007. - V. 7. - 3. - p. 241-60.
47. Orlov Y.L. Z.J.T., Chen J., Shahab A., Kuznetsov V.A., *APMA Database for Affymetrix target sequences mapping, quality assessment and expression data mining*, in *Pattern Recognition in Bioinformatics: second IAPR international workshop, PRIB 2007*, B.S. J.C. Ragapakse, G. Volkert, Editor. 2007, Springer-Verlag: Berlin-Heidelberg. p. 166-177.
48. 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. *Nucleic Acids Res* - 2010. - V. 38. - 2. - p. 534-47.
49. Орлов Ю.Л. Ефимов В.М., Орлова Н.Г. Статистические оценки экспрессии мобильных элементов в геноме человека на основе клинических данных экспрессионных микрочипов. *Вавиловский журнал генетики и селекции* - 2011. - V. 15. - 2. - p. 327-339.
50. Орлов Ю.Л. Левицкий В.Г., Смирнова О.Г., Подколодная О.А., Хлебодарова Т.М., Колчанов Н.А. Статистический анализ последовательностей ДНК, содержащих сайты формирования нуклеосом. *Биофизика* - 2006. - V. 51. - p. 608-614.
51. Goh W.S., Orlov Y., Li J., et al. Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. *PLoS Comput Biol* - 2010. - V. 6. - 1. - p. e1000649.
52. Heng J.C., Orlov Y.L. and Ng H.H. Transcription factors for the modulation of pluripotency and reprogramming. *Cold Spring Harb Symp Quant Biol* - 2010. - V. 75. - p. 237-44.
53. Кожевникова О.С. Мартыщенко М.К., Генаев М.К., Корболина М.К., Муралева Н.А., Колосова Н.А., Орлов Ю.Л. RatDNA: база данных микрочиповых исследований на крысах для генов, ассоциированных с заболеваниями старения. *Вавиловский журнал генетики и селекции* - 2012. - V. 16. - 4/1. - p. 756-765.
54. Lee K.L., Lim S.K., Orlov Y.L., et al. Graded Nodal/Activin signaling titrates conversion of quantitative phospho-Smad2 levels into qualitative embryonic stem cell fate decisions. *PLoS Genet* - 2011. - V. 7. - 6. - p. e1002130.

55. Орлов Ю.Л. Компьютерное исследование регуляции транскрипции генов эукариот с помощью данных экспериментов секвенирования и иммунопреципитации хроматина. *Вавиловский журнал генетики и селекции* - 2014. -Т. 18, - 1, С. 193-206.
56. Guo X., Popadin K.Y., Markuzon N., et al. Repeats, longevity and the sources of mtDNA deletions: evidence from 'deletional spectra'. *Trends Genet* - 2010. - V. 26. - 8. - p. 340-3.
57. Orlov Y.L., Te Boekhorst R. and Abnizova, II. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. *J Bioinform Comput Biol* - 2006. - V. 4. - 2. - p. 523-36.
58. Путта П. Орлов Ю.Л., Подколодный Н.Л., Митра Ч.К. Относительно консервативные общие короткие последовательности в сайтах связывания транскрипционных факторов и миРНК. *Вавиловский журнал генетики и селекции* - 2011. - V. 15. - 4. - p. 750-756.
59. Воробьева Н.В. Билтуева Л.С., Орлов Ю.Л., Графодатский А.С., Колчанов Н.А. Интерстициальные теломерные повторы, как маркеры эволюционных преобразований кариотипа млекопитающих: хромосома 2 человека. *Биофизика* - 2006. - V. 51. - p. 602-7.
60. Медведева И.В. Вишневский О.В., Сафронова Н.С., Кожевникова О.С., Генаев М.А., Кочетов А.В., Афонников Д.А., Орлов Ю.Л. Компьютерный анализ данных экспрессии генов в клетках мозга, полученных с помощью микрочипов и высокопроизводительного секвенирования. *Вавиловский журнал генетики и селекции* - 2013. - V. 17. - 4/1. - p. 629-638.
61. Orlov Y.L. D.O., Yuan C.H., Afonnikov D.A., Zhu Y., Chen M. Integrative computer analysis of antisense transcripts and miRNA targets in plant genomes. *Journal of Stress Physiology & Biochemistry* - 2012. - V. 8. - 3. - p. S7.
62. Matushkin Y.G., Levitsky V.G., Orlov Y.L., et al. Translation efficiency in yeasts correlates with nucleosome formation in promoters. *J Biomol Struct Dyn* - 2013. - V. 31. - 1. - p. 96-102.
63. Матушкин Ю.Г. Левицкий В.Г., Соколов В.С., Лихошвай В.А., Орлов Ю.Л. Эффективность элонгации генов дрожжей коррелирует с плотностью нуклеосомной упаковки в 5'-нетранслируемом районе. *Математическая биология и биоинформатика* - 2013. - V. 8. - 1. - p. 248-257.
64. Баттулин Н.Р. Ф.В.С., Орлов Ю.Л., Мензоров А.Г., Афонников Д.А., Серов О.Л. 3С-методы в исследованиях пространственной организации генома. *Вавиловский журнал генетики и селекции* - 2012. - V. 16. - 4/2. - p. 872-878.
65. Muers M. Functional genomics: the modENCODE guide to the genome. *Nat Rev Genet* - 2011. - V. 12. - 2. - p. 80.
66. Gerstein M.B., Lu Z.J., Van Nostrand E.L., et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* - 2010. - V. 330. - 6012. - p. 1775-87.
67. Benson D.A., Clark K., Karsch-Mizrachi I., et al. GenBank. *Nucleic Acids Res* - 2014. - V. 42. - 1. - p. D32-7.
68. Adhikary S. and Eilers M. Transcriptional regulation and transformation by Myc proteins. *Nat Rev Mol Cell Biol* - 2005. - V. 6. - 8. - p. 635-45.
69. Altschul S., Demchak B., Durbin R., et al. The anatomy of successful computational biology software. *Nat Biotechnol* - 2013. - V. 31. - 10. - p. 894-7.
70. Rothberg J.M., Hinz W., Rearick T.M., et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* - 2011. - V. 475. - 7356. - p. 348-52.
71. Kolchanov N.A. L.H.A., ed. *Computer Analysis of Genetic Macromolecules: Structure, Function and Evolution*. 1994, World Scientific Pub. co.: Singapore, New Jersey, London, Hong Kong. 556.

72. Medvedeva I., Demenkov P., Kolchanov N., et al. SitEx: a computer system for analysis of projections of protein functional sites on eukaryotic genes. *Nucleic Acids Res* - 2012. - V. 40. - Database issue. - p. D278-83.
73. Lewin B., *Genes VII*. 2000, Oxford: Oxford University Press. 990.
74. Toth G., Gaspari Z. and Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* - 2000. - V. 10. - 7. - p. 967-81.
75. Jurka J., Kapitonov V.V., Pavlicek A., et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* - 2005. - V. 110. - 1-4. - p. 462-7.
76. Jurka J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* - 2000. - V. 16. - 9. - p. 418-20.
77. Smit A.F.A. H.R., Green P., *RepeatMasker Open-3.0*. 1996-2010.
78. Altschul S.F., Gish W., Miller W., et al. Basic local alignment search tool. *J Mol Biol* - 1990. - V. 215. - 3. - p. 403-10.
79. Needleman S.B. and Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* - 1970. - V. 48. - 3. - p. 443-53.
80. Bladon P. A simple method for aligning many protein sequences. *J Chem Inf Comput Sci* - 2001. - V. 41. - 2. - p. 278-80.
81. Karlin S., Ghandour G., Ost F., et al. New approaches for computer analysis of nucleic acid sequences. *Proc Natl Acad Sci U S A* - 1983. - V. 80. - 18. - p. 5660-4.
82. Kurtz S., Choudhuri J.V., Ohlebusch E., et al. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res* - 2001. - V. 29. - 22. - p. 4633-42.
83. Soares I., Goios A. and Amorim A. Sequence comparison alignment-free approach based on suffix tree and L-words frequency. *ScientificWorldJournal* - 2012. - V. 2012. - p. 450124.
84. Apostolico A., Bock M.E., Lonardi S., et al. Efficient detection of unusual words. *J Comput Biol* - 2000. - V. 7. - 1-2. - p. 71-94.
85. Pearson W.R. and Lipman D.J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* - 1988. - V. 85. - 8. - p. 2444-8.
86. Kent W.J. BLAT--the BLAST-like alignment tool. *Genome Res* - 2002. - V. 12. - 4. - p. 656-64.
87. Lee H. and Schatz M.C. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* - 2012. - V. 28. - 16. - p. 2097-105.
88. Sneath P.H.A. S.R.R., *Numerical Taxonomy. The principles and practice of numerical classification*. 1973, San Francisco: W.H. Freeman and Co. 573.
89. Nei M. Phylogenetic analysis in molecular evolutionary genetics. *Annu Rev Genet* - 1996. - V. 30. - p. 371-403.
90. Морозов П.С., *Новые методы оценки параметров эволюционного процесса при филогенетическом анализе.* 2000, ИЦиГ СО РАН: Новосибирск. p. 157.
91. Kumar S., Tamura K., Jakobsen I.B., et al. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* - 2001. - V. 17. - 12. - p. 1244-5.
92. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* - 1997. - V. 13. - 5. - p. 555-6.
93. Zharkikh A. and Li W.H. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Mol Biol Evol* - 1992. - V. 9. - 6. - p. 1119-47.
94. Zharkikh A.A., Rzhetsky A., Morosov P.S., et al. VOSTORG: a package of microcomputer programs for sequence analysis and construction of phylogenetic trees. *Gene* - 1991. - V. 101. - 2. - p. 251-4.

95. Felsenstein J. An alternating least squares approach to inferring phylogenies from pairwise distances. *Syst Biol* - 1997. - V. 46. - 1. - p. 101-11.
96. Gunbin K.V., Suslov V.V., Genaev M.A., et al. Computer System for Analysis of Molecular Evolution Modes (SAMEM): analysis of molecular evolution modes at deep inner branches of the phylogenetic tree. *In Silico Biol* - 2011-2012. - V. 11. - 3-4. - p. 109-23.
97. Bishop M.J., ed. *Guide to human genome computing*. 1998, Academic Press: London 306.
98. Fisher R.A. On the Interpretation of x2 from Contingency Tables, and the Calculation of P. *J. Roy. Statistical Society* - 1922. - V. 85. - 1. - p. 87-94.
99. Ernst J., Plasterer H.L., Simon I., et al. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res* - 2010. - V. 20. - 4. - p. 526-36.
100. Matsumoto M. N.T. Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation* - 1998. - V. 8. - p. 3-30.
101. Click T.H., Liu A. and Kaminski G.A. Quality of random number generators significantly affects results of Monte Carlo simulations for organic and biological systems. *J Comput Chem* - 2011. - V. 32. - 3. - p. 513-24.
102. Ратнер В.А. Генетический язык; грамматика, семантика, эволюция. *Генетика* - 1993. - V. 29. - 5. - p. 709-719.
103. Ратнер В.А. *Молекулярно-генетические системы управления*. 1975, Новосибирск: Наука. Сиб. отделение. 257.
104. Trifonov E.N. The multiple codes of nucleotide sequences. *Bull Math Biol* - 1989. - V. 51. - 4. - p. 417-32.
105. Troyanskaya O.G., Arbell O., Koren Y., et al. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. *Bioinformatics* - 2002. - V. 18. - 5. - p. 679-88.
106. Wan H., Li L., Federhen S., et al. Discovering simple regions in biological sequences associated with scoring schemes. *J Comput Biol* - 2003. - V. 10. - 2. - p. 171-85.
107. Osmanbeyoglu H.U. and Ganapathiraju M.K. N-gram analysis of 970 microbial organisms reveals presence of biological language models. *BMC Bioinformatics* - 2011. - V. 12. - p. 12.
108. Popov O., Segal D.M. and Trifonov E.N. Linguistic complexity of protein sequences as compared to texts of human languages. *Biosystems* - 1996. - V. 38. - 1. - p. 65-74.
109. Zhurkin V.B. Periodicity in DNA primary structure is defined by secondary structure of the coded protein. *Nucleic Acids Res* - 1981. - V. 9. - 8. - p. 1963-71.
110. Trifonov E.N. Thirty years of multiple sequence codes. *Genomics Proteomics Bioinformatics* - 2011. - V. 9. - 1-2. - p. 1-6.
111. Cohanim A.B. and Haran T.E. The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes. *Nucleic Acids Res* - 2009. - V. 37. - 19. - p. 6466-76.
112. Parker S.C. and Tullius T.D. DNA shape, genetic codes, and evolution. *Curr Opin Struct Biol* - 2011. - V. 21. - 3. - p. 342-7.
113. Baisnee P.F., Baldi P., Brunak S., et al. Flexibility of the genetic code with respect to DNA structure. *Bioinformatics* - 2001. - V. 17. - 3. - p. 237-48.
114. Orlov Y.L. and Potapov V.N. Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic Acids Res* - 2004. - V. 32. - Web Server issue. - p. W628-33.
115. Гусев В.Д. Куличков В.А., Чупахина О.М. Анализ сложности геномов. Мера сложности и классификация выявленных структурных особенностей. *Молекулярная биология* - 1991. - V. 25. - p. 825-834.

116. Gusev V.D., Nemytikova L.A. and Chuzhanova N.A. On the complexity measures of genetic sequences. *Bioinformatics* - 1999. - V. 15. - 12. - p. 994-9.
117. Wootton J.C. and Federhen S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* - 1996. - V. 266. - p. 554-71.
118. Trifonov E.N., *Making Sense of the Human Genome*, in *Structure & Methods*, S.M.H. Sarma R.H., Editor. 1990, Adenine Press: Albany. p. 69-77.
119. Core L.J., Waterfall J.J. and Lis J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* - 2008. - V. 322. - 5909. - p. 1845-8.
120. Rhee H.S. and Pugh B.F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* - 2011. - V. 147. - 6. - p. 1408-19.
121. Kolchanov N.A., Ignatieva E.V., Ananko E.A., et al. Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res* - 2002. - V. 30. - 1. - p. 312-7.
122. Лихошвай В.А. Матушкин Ю.Г. Предсказание эффективности генной экспрессии по нуклеотидному составу. *Молекулярная биология* - 2000. - V. 34. - 3. - p. 406-412.
123. Kadener S., Fededa J.P., Rosbash M., et al. Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. *Proc Natl Acad Sci U S A* - 2002. - V. 99. - 12. - p. 8185-90.
124. Wray G.A., Hahn M.W., Abouheif E., et al. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* - 2003. - V. 20. - 9. - p. 1377-419.
125. Албертс Б. Брей Д., Льюис Дж., Рэфф М., Робертс К., Уотсон Дж. Д., *Молекулярная биология клетки*. Vol. 2. 1993, Москва: Мир. 539.
126. Nikolov D.B. and Burley S.K. RNA polymerase II transcription initiation: a structural view. *Proc Natl Acad Sci U S A* - 1997. - V. 94. - 1. - p. 15-22.
127. Emerson B.M. Specificity of gene regulation. *Cell* - 2002. - V. 109. - 3. - p. 267-70.
128. Патрушев Л.И., *Экспрессия генов*. 2000, Москва: Наука. 830.
129. Matys V., Kel-Margoulis O.V., Fricke E., et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* - 2006. - V. 34. - Database issue. - p. D108-10.
130. Ham J., Steger G. and Yaniv M. How do eukaryotic activator proteins stimulate the rate of transcription by RNA polymerase II? *FEBS Lett* - 1992. - V. 307. - 1. - p. 81-6.
131. Manley J.L., Um M., Li C., et al. Mechanisms of transcriptional activation and repression can both involve TFIID. *Philos Trans R Soc Lond B Biol Sci* - 1996. - V. 351. - 1339. - p. 517-26.
132. Rhee H.S. and Pugh B.F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* - 2012. - V. 483. - 7389. - p. 295-301.
133. Tjian R. The biochemistry of transcription in eukaryotes: a paradigm for multisubunit regulatory complexes. *Philos Trans R Soc Lond B Biol Sci* - 1996. - V. 351. - 1339. - p. 491-9.
134. Thomas M.C. and Chiang C.M. E6 oncoprotein represses p53-dependent gene activation via inhibition of protein acetylation independently of inducing p53 degradation. *Mol Cell* - 2005. - V. 17. - 2. - p. 251-64.
135. Myers L.C. and Kornberg R.D. Mediator of transcriptional regulation. *Annu Rev Biochem* - 2000. - V. 69. - p. 729-49.
136. Kim Y.J., Bjorklund S., Li Y., et al. A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell* - 1994. - V. 77. - 4. - p. 599-608.
137. Malik S. and Roeder R.G. Dynamic regulation of pol II transcription by the mammalian Mediator complex. *Trends Biochem Sci* - 2005. - V. 30. - 5. - p. 256-63.

138. Жимулев И.Ф., *Общая и молекулярная генетика*. Учеб. пособие.-2-е изд. 2003, Новосибирск: Сиб.унив. изд-во. 479.
139. Versteeg R., van Schaik B.D., van Batenburg M.F., et al. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* - 2003. - V. 13. - 9. - p. 1998-2004.
140. Eisenberg E. and Levanon E.Y. Human housekeeping genes are compact. *Trends Genet* - 2003. - V. 19. - 7. - p. 362-5.
141. Taylor J. Clues to function in gene deserts. *Trends Biotechnol* - 2005. - V. 23. - 6. - p. 269-71.
142. Hoey T., Dynlacht B.D., Peterson M.G., et al. Isolation and characterization of the Drosophila gene encoding the TATA box binding protein, TFIID. *Cell* - 1990. - V. 61. - 7. - p. 1179-86.
143. Ossipow V., Tassan J.P., Nigg E.A., et al. A mammalian RNA polymerase II holoenzyme containing all components required for promoter-specific transcription initiation. *Cell* - 1995. - V. 83. - 1. - p. 137-46.
144. Hatfield G.W., Hung S.P. and Baldi P. Differential analysis of DNA microarray gene expression data. *Mol Microbiol* - 2003. - V. 47. - 4. - p. 871-7.
145. Stollberg J., Urschitz J., Urban Z., et al. A quantitative evaluation of SAGE. *Genome Res* - 2000. - V. 10. - 8. - p. 1241-8.
146. Lu J., Lal A., Merriman B., et al. A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. *Genomics* - 2004. - V. 84. - 4. - p. 631-6.
147. Su A.I., Cooke M.P., Ching K.A., et al. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* - 2002. - V. 99. - 7. - p. 4465-70.
148. Liu G., Loraine A.E., Shigeta R., et al. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res* - 2003. - V. 31. - 1. - p. 82-6.
149. Shames D.S., Girard L., Gao B., et al. A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. *PLoS Med* - 2006. - V. 3. - 12. - p. e486.
150. Dai M., Wang P., Boyd A.D., et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* - 2005. - V. 33. - 20. - p. e175.
151. Harbig J., Sprinkle R. and Enkemann S.A. A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res* - 2005. - V. 33. - 3. - p. e31.
152. Sela N., Mersch B., Hotz-Wagenblatt A., et al. Characteristics of transposable element exonization within human and mouse. *PLoS One* - 2010. - V. 5. - 6. - p. e10907.
153. Tusher V.G., Tibshirani R. and Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* - 2001. - V. 98. - 9. - p. 5116-21.
154. Affymetrix, *MAS 5.0 algorithm. Statistical Algorithms Description Document.* . 2002, Affymetrix, Inc. Santa Clara, CA.
155. Chudin E., Walker R., Kosaka A., et al. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol* - 2002. - V. 3. - 1. - p. RESEARCH0005.
156. Gautier L., Cope L., Bolstad B.M., et al. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* - 2004. - V. 20. - 3. - p. 307-15.
157. Okoniewski M.J. and Miller C.J. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics* - 2006. - V. 7. - p. 276.

158. Stalteri M.A. and Harrison A.P. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics* - 2007. - V. 8. - p. 13.
159. Fasold M., Stadler P.F. and Binder H. G-stack modulated probe intensities on expression arrays - sequence corrections and signal calibration. *BMC Bioinformatics* - 2010. - V. 11. - p. 207.
160. Nellaker C., Li F., Uhrzander F., et al. Expression profiling of repetitive elements by melting temperature analysis: variation in HERV-W gag expression across human individuals and tissues. *BMC Genomics* - 2009. - V. 10. - p. 532.
161. Karlsson H., Bachmann S., Schroder J., et al. Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. *Proc Natl Acad Sci U S A* - 2001. - V. 98. - 8. - p. 4634-9.
162. Frank O., Verbeke C., Schwarz N., et al. Variable transcriptional activity of endogenous retroviruses in human breast cancer. *J Virol* - 2008. - V. 82. - 4. - p. 1808-18.
163. Свердлов Е.Д., *Очерки структурной молекулярной генетики*. Взгляд на жизнь через окно генома. Vol. 1. 2009, Москва: Наука. 525.
164. Shenk T. Transcriptional control regions: nucleotide sequence requirements for initiation by RNA polymerase II and III. *Curr Top Microbiol Immunol* - 1981. - V. 93. - p. 25-46.
165. Arnone M.I. and Davidson E.H. The hardwiring of development: organization and function of genomic regulatory systems. *Development* - 1997. - V. 124. - 10. - p. 1851-64.
166. Bucher P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* - 1990. - V. 212. - 4. - p. 563-78.
167. 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. - 6. - p. 424-36.
168. Kadonaga J.T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol* - 2012. - V. 1. - 1. - p. 40-51.
169. Ohler U., Liao G.C., Niemann H., et al. Computational analysis of core promoters in the Drosophila genome. *Genome Biol* - 2002. - V. 3. - 12. - p. RESEARCH0087.
170. Притчард Д.Дж. К.Б.Р., *Наглядная медицинская генетика*. 2009: ГЭОТАР-Медиа. 200.
171. Arkhipova I.R. Promoter elements in Drosophila melanogaster revealed by sequence analysis. *Genetics* - 1995. - V. 139. - 3. - p. 1359-69.
172. Burke T.W. and Kadonaga J.T. The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev* - 1997. - V. 11. - 22. - p. 3020-31.
173. Dieci G., Bosio M.C., Fermi B., et al. Transcription reinitiation by RNA polymerase III. *Biochim Biophys Acta* - 2013. - V. 1829. - 3-4. - p. 331-41.
174. Goodfellow S.J. and Zomerdijk J.C. Basic mechanisms in RNA polymerase I transcription of the ribosomal RNA genes. *Subcell Biochem* - 2012. - V. 61. - p. 211-36.
175. de Laat W. and Grosveld F. Spatial organization of gene expression: the active chromatin hub. *Chromosome Res* - 2003. - V. 11. - 5. - p. 447-59.
176. Khoury G. and Gruss P. Enhancer elements. *Cell* - 1983. - V. 33. - 2. - p. 313-4.
177. Herr W. and Clarke J. The SV40 enhancer is composed of multiple functional elements that can compensate for one another. *Cell* - 1986. - V. 45. - 3. - p. 461-70.
178. Kim A. and Dean A. Chromatin loop formation in the beta-globin locus and its role in globin gene transcription. *Mol Cells* - 2012. - V. 34. - 1. - p. 1-5.

179. Amano T., Sagai T., Tanabe H., et al. Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev Cell* - 2009. - V. 16. - 1. - p. 47-57.
180. Lettice L.A., Horikoshi T., Heaney S.J., et al. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci U S A* - 2002. - V. 99. - 11. - p. 7548-53.
181. Ong C.T. and Corces V.G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* - 2011. - V. 12. - 4. - p. 283-93.
182. 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. - 7243. - p. 108-12.
183. Visel A., Rubin E.M. and Pennacchio L.A. Genomic views of distant-acting enhancers. *Nature* - 2009. - V. 461. - 7261. - p. 199-205.
184. Spellman P.T. and Rubin G.M. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol* - 2002. - V. 1. - 1. - p. 5.
185. Singer G.A., Lloyd A.T., Huminiecki L.B., et al. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol Biol Evol* - 2005. - V. 22. - 3. - p. 767-75.
186. Stolc V., Gauhar Z., Mason C., et al. A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science* - 2004. - V. 306. - 5696. - p. 655-60.
187. Semon M. and Duret L. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol Biol Evol* - 2006. - V. 23. - 9. - p. 1715-23.
188. Phillips J.E. and Corces V.G. CTCF: master weaver of the genome. *Cell* - 2009. - V. 137. - 7. - p. 1194-211.
189. Pati D., Zhang N. and Plon S.E. Linking sister chromatid cohesion and apoptosis: role of Rad21. *Mol Cell Biol* - 2002. - V. 22. - 23. - p. 8267-77.
190. Euskirchen G.M., Auerbach R.K., Davidov E., et al. Diverse roles and interactions of the SWI/SNF chromatin remodeling complex revealed using global approaches. *PLoS Genet* - 2011. - V. 7. - 3. - p. e1002008.
191. Ni Z., Abou El Hassan M., Xu Z., et al. The chromatin-remodeling enzyme BRG1 coordinates CIITA induction through many interdependent distal enhancers. *Nat Immunol* - 2008. - V. 9. - 7. - p. 785-93.
192. Werner M.H. and Burley S.K. Architectural transcription factors: proteins that remodel DNA. *Cell* - 1997. - V. 88. - 6. - p. 733-6.
193. Panne D. The enhanceosome. *Curr Opin Struct Biol* - 2008. - V. 18. - 2. - p. 236-42.
194. Thanos D. and Maniatis T. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* - 1995. - V. 83. - 7. - p. 1091-100.
195. Arnosti D.N. and Kulkarni M.M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* - 2005. - V. 94. - 5. - p. 890-8.
196. Panne D., Maniatis T. and Harrison S.C. An atomic model of the interferon-beta enhanceosome. *Cell* - 2007. - V. 129. - 6. - p. 1111-23.
197. Blackwood E.M. and Kadonaga J.T. Going the distance: a current view of enhancer action. *Science* - 1998. - V. 281. - 5373. - p. 60-3.
198. Taher L., Smith R.P., Kim M.J., et al. Sequence signatures extracted from proximal promoters can be used to predict distal enhancers. *Genome Biol* - 2013. - V. 14. - 10. - p. R117.
199. Kel O.V., Romaschenko A.G., Kel A.E., et al. A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res* - 1995. - V. 23. - 20. - p. 4097-103.
200. Kolchanov N.A., Podkolodnaya O.A., Ananko E.A., et al. Transcription regulatory regions database (TRRD): its status in 2000. *Nucleic Acids Res* - 2000. - V. 28. - 1. - p. 298-301.

201. Lee C., Atanelov L., Modrek B., et al. ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res* - 2003. - V. 31. - 1. - p. 101-5.
202. Gelfand M.S., Dubchak I., Dralyuk I., et al. ASDB: database of alternatively spliced genes. *Nucleic Acids Res* - 1999. - V. 27. - 1. - p. 301-2.
203. Chew J.L., Loh Y.H., Zhang W., et al. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol Cell Biol* - 2005. - V. 25. - 14. - p. 6031-46.
204. Vanyushin B.F., Tkacheva S.G. and Belozersky A.N. Rare bases in animal DNA. *Nature* - 1970. - V. 225. - 5236. - p. 948-9.
205. Jiang Y., Liu S., Chen X., et al. Genome-wide distribution of DNA methylation and DNA demethylation and related chromatin regulators in cancer. *Biochim Biophys Acta* - 2013. - V. 1835. - 2. - p. 155-63.
206. Baylin S.B., Herman J.G., Graff J.R., et al. Alterations in DNA methylation: a fundamental aspect of neoplasia. *Adv Cancer Res* - 1998. - V. 72. - p. 141-96.
207. Vanyushin B.F. A view of an elemental naturalist at the DNA world (base composition, sequences, methylation). *Biochemistry (Mosc)* - 2007. - V. 72. - 12. - p. 1289-98.
208. Hotchkiss R.D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J Biol Chem* - 1948. - V. 175. - 1. - p. 315-32.
209. Wyatt G.R. Recognition and estimation of 5-methylcytosine in nucleic acids. *Biochem J* - 1951. - V. 48. - 5. - p. 581-4.
210. Vinson C. and Chatterjee R. CG methylation. *Epigenomics* - 2012. - V. 4. - 6. - p. 655-63.
211. Jones P.L. and Wolffe A.P. Relationships between chromatin organization and DNA methylation in determining gene expression. *Semin Cancer Biol* - 1999. - V. 9. - 5. - p. 339-47.
212. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* - 2013. - V. 14. - 10. - R115.
213. Jones P.L., Veenstra G.J., Wade P.A., et al. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet* - 1998. - V. 19. - 2. - p. 187-91.
214. Issa J.P., Ottaviano Y.L., Celano P., et al. Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. *Nat Genet* - 1994. - V. 7. - 4. - p. 536-40.
215. Ahuja N., Li Q., Mohan A.L., et al. Aging and DNA methylation in colorectal mucosa and cancer. *Cancer Res* - 1998. - V. 58. - 23. - p. 5489-94.
216. Lapidus R.G., Ferguson A.T., Ottaviano Y.L., et al. Methylation of estrogen and progesterone receptor gene 5' CpG islands correlates with lack of estrogen and progesterone receptor gene expression in breast tumors. *Clin Cancer Res* - 1996. - V. 2. - 5. - p. 805-10.
217. Kel A.E., Gossling E., Reuter I., et al. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* - 2003. - V. 31. - 13. - p. 3576-9.
218. Sandelin A., Alkema W., Engstrom P., et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* - 2004. - V. 32. - Database issue. - p. D91-4.
219. Zhao Y., Granas D. and Stormo G.D. Inferring binding energies from selected binding sites. *PLoS Comput Biol* - 2009. - V. 5. - 12. - p. e1000590.
220. Foat B.C., Morozov A.V. and Bussemaker H.J. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* - 2006. - V. 22. - 14. - p. e141-9.
221. Поздняков М.А. Витяев Е.Е., Ананько Е.А., Игнатьева Е.В.,. Подколодная О.А, Подколодный Н.Л., Лаврюшев С.В., Колчанов Н.А. Сравнительный анализ

методов распознавания потенциальных сайтов связывания транскрипционных факторов. *Молекулярная биология* - 2001. - V. 35. - p. 961-969.

222. Kulakovskiy I.V., Medvedeva Y.A., Schaefer U., et al. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res* - 2013. - V. 41. - Database issue. - p. D195-202.
223. Kondrakhin Y.V., Kel A.E., Kolchanov N.A., et al. Eukaryotic promoter recognition by binding sites for transcription factors. *Comput Appl Biosci* - 1995. - V. 11. - 5. - p. 477-88.
224. Fickett J.W. and Hatzigeorgiou A.G. Eukaryotic promoter recognition. *Genome Res* - 1997. - V. 7. - 9. - p. 861-78.
225. Werner T. The state of the art of mammalian promoter recognition. *Brief Bioinform* - 2003. - V. 4. - 1. - p. 22-30.
226. Koch C., Moll T., Neuberg M., et al. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science* - 1993. - V. 261. - 5128. - p. 1551-7.
227. Taylor I.A., McIntosh P.B., Pala P., et al. Characterization of the DNA-binding domains from the yeast cell-cycle transcription factors Mbp1 and Swi4. *Biochemistry (Mosc)* - 2000. - V. 39. - 14. - p. 3943-54.
228. Bork P., Dandekar T., Diaz-Lazcoz Y., et al. Predicting function: from genes to genomes and back. *J Mol Biol* - 1998. - V. 283. - 4. - p. 707-25.
229. Tiwari S., Ramachandran S., Bhattacharya A., et al. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput Appl Biosci* - 1997. - V. 13. - 3. - p. 263-70.
230. Pavesi G., Mauri G. and Pesole G. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* - 2001. - V. 17 Suppl 1. - p. S207-14.
231. van Helden J. Regulatory sequence analysis tools. *Nucleic Acids Res* - 2003. - V. 31. - 13. - p. 3593-6.
232. Zhou Q. and Liu J.S. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* - 2004. - V. 20. - 6. - p. 909-16.
233. Klingenhoff A., Frech K., Quandt K., et al. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* - 1999. - V. 15. - 3. - p. 180-6.
234. Thijs G., Lescot M., Marchal K., et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* - 2001. - V. 17. - 12. - p. 1113-22.
235. Huang H.D., Horng J.T., Sun Y.M., et al. Identifying transcriptional regulatory sites in the human genome using an integrated system. *Nucleic Acids Res* - 2004. - V. 32. - 6. - p. 1948-56.
236. Шматков А.М. Исследование локальных корреляций в генах бактерий для использования в рамках скрытых марковских моделей. *Молекулярная биология* - 2000. - V. 34. - 5. - p. 868-874.
237. Madera M. and Gough J. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res* - 2002. - V. 30. - 19. - p. 4321-8.
238. Pachter L., Alexandersson M. and Cawley S. Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J Comput Biol* - 2002. - V. 9. - 2. - p. 389-99.
239. Viterbi A.J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Informat. Theory* - 1967. - V. V.IT-13. - p. 260-269.
240. Gadiraju S., Vyhlidal C.A., Leeder J.S., et al. Genome-wide prediction, display and refinement of binding sites with information theory-based models. *BMC Bioinformatics* - 2003. - V. 4. - p. 38.
241. Teufel A., Krupp M., Weinmann A., et al. Current bioinformatics tools in genomic biomedical research (Review). *Int J Mol Med* - 2006. - V. 17. - 6. - p. 967-73.

242. Bajic V.B., Seah S.H., Chong A., et al. Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. *Bioinformatics* - 2002. - V. 18. - 1. - p. 198-9.
243. Berg O.G. and von Hippel P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* - 1987. - V. 193. - 4. - p. 723-50.
244. Пономаренко М.П., Пономаренко Ю.В., Кель А.Э., Колчанов Н.А. Компьютерный анализ конформационных свойств ТАТА-бокса в промоторах эукариот. *Молекулярная биология* - 1997. - V. 31. - p. 733-740.
245. Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., et al. Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis. *Bioinformatics* - 2001. - V. 17. - 11. - p. 998-1010.
246. Frisch M., Frech K., Klingenhoff A., et al. In silico prediction of scaffold/matrix attachment regions in large genomic sequences. *Genome Res* - 2002. - V. 12. - 2. - p. 349-54.
247. Zhang M.Q. Discriminant analysis and its application in DNA sequence motif recognition. *Brief Bioinform* - 2000. - V. 1. - 4. - p. 331-42.
248. Hutchinson G.B. The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput Appl Biosci* - 1996. - V. 12. - 5. - p. 391-8.
249. Solovyev V. and Salamov A. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Proc Int Conf Intell Syst Mol Biol* - 1997. - V. 5. - p. 294-302.
250. Shahmuradov I.A., Gammerman A.J., Hancock J.M., et al. PlantProm: a database of plant promoter sequences. *Nucleic Acids Res* - 2003. - V. 31. - 1. - p. 114-7.
251. Krebs J.E. Goldstein E.S., Kilpatrick S.T., *Lewin's GENES XI (11th ed)*. 2012: Jones & Bartlett Publishers. 939.
252. Khorasanizadeh S. The nucleosome: from genomic organization to genomic regulation. *Cell* - 2004. - V. 116. - 2. - p. 259-72.
253. Widom J. Role of DNA sequence in nucleosome stability and dynamics. *Q Rev Biophys* - 2001. - V. 34. - 3. - p. 269-324.
254. Kiyama R. and Trifonov E.N. What positions nucleosomes?--A model. *FEBS Lett* - 2002. - V. 523. - 1-3. - p. 7-11.
255. Levitsky V.G., Ponomarenko M.P., Ponomarenko J.V., et al. Nucleosomal DNA property database. *Bioinformatics* - 1999. - V. 15. - 7-8. - p. 582-92.
256. Levitsky V.G. RECON: a program for prediction of nucleosome formation potential. *Nucleic Acids Res* - 2004. - V. 32. - Web Server issue. - p. W346-9.
257. Levitsky V.G., Babenko V.N. and Vershinin A.V. The roles of the monomer length and nucleotide context of plant tandem repeats in nucleosome positioning. *J Biomol Struct Dyn* - 2014. - V. 32. - 1. - p. 115-26.
258. Soutoglou E. and Talianidis I. Coordination of PIC assembly and chromatin remodeling during differentiation-induced gene activation. *Science* - 2002. - V. 295. - 5561. - p. 1901-4.
259. Ioshikhes I., Bolshoy A., Derenshteyn K., et al. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J Mol Biol* - 1996. - V. 262. - 2. - p. 129-39.
260. Segal E., Fondufe-Mittendorf Y., Chen L., et al. A genomic code for nucleosome positioning. *Nature* - 2006. - V. 442. - 7104. - p. 772-8.
261. Segal E. and Widom J. What controls nucleosome positions? *Trends Genet* - 2009. - V. 25. - 8. - p. 335-43.
262. Segal E. and Widom J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol* - 2009. - V. 19. - 1. - p. 65-71.

263. Reynolds S.M., Bilmes J.A. and Noble W.S. Learning a weighted sequence model of the nucleosome core and linker yields more accurate predictions in Saccharomyces cerevisiae and Homo sapiens. *PLoS Comput Biol* - 2010. - V. 6. - 7. - p. e1000834.
264. Kaplan N., Moore I.K., Fondufe-Mittendorf Y., et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* - 2009. - V. 458. - 7236. - p. 362-6.
265. Tanaka Y. and Nakai K. An assessment of prediction algorithms for nucleosome positioning. *Genome Inform* - 2009. - V. 23. - 1. - p. 169-78.
266. Peckham H.E., Thurman R.E., Fu Y., et al. Nucleosome positioning signals in genomic DNA. *Genome Res* - 2007. - V. 17. - 8. - p. 1170-7.
267. Gupta S., Dennis J., Thurman R.E., et al. Predicting human nucleosome occupancy from primary sequence. *PLoS Comput Biol* - 2008. - V. 4. - 8. - p. e1000134.
268. Ioshikhes I. and Trifonov E.N. Nucleosomal DNA sequence database. *Nucleic Acids Res* - 1993. - V. 21. - 21. - p. 4857-9.
269. Shendure J. and Ji H. Next-generation DNA sequencing. *Nat Biotechnol* - 2008. - V. 26. - 10. - p. 1135-45.
270. Taverna S.D., Ueberheide B.M., Liu Y., et al. Long-distance combinatorial linkage between methylation and acetylation on histone H3 N termini. *Proc Natl Acad Sci U S A* - 2007. - V. 104. - 7. - p. 2086-91.
271. Giresi P.G., Kim J., McDaniell R.M., et al. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* - 2007. - V. 17. - 6. - p. 877-85.
272. Gilbert N. and Ramsahoye B. The relationship between chromatin structure and transcriptional activity in mammalian genomes. *Brief Funct Genomic Proteomic* - 2005. - V. 4. - 2. - p. 129-42.
273. Solomon M.J., Larsen P.L. and Varshavsky A. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* - 1988. - V. 53. - 6. - p. 937-47.
274. Ren B., Robert F., Wyrick J.J., et al. Genome-wide location and function of DNA binding proteins. *Science* - 2000. - V. 290. - 5500. - p. 2306-9.
275. Lee C.K., Shibata Y., Rao B., et al. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet* - 2004. - V. 36. - 8. - p. 900-5.
276. Yuan G.C., Liu Y.J., Dion M.F., et al. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* - 2005. - V. 309. - 5734. - p. 626-30.
277. Pokholok D.K., Harbison C.T., Levine S., et al. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* - 2005. - V. 122. - 4. - p. 517-27.
278. Harbison C.T., Gordon D.B., Lee T.I., et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* - 2004. - V. 431. - 7004. - p. 99-104.
279. MacIsaac K.D., Wang T., Gordon D.B., et al. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* - 2006. - V. 7. - p. 113.
280. Johnson R., Teh C.H., Kunarso G., et al. REST regulates distinct transcriptional networks in embryonic and neural stem cells. *PLoS Biol* - 2008. - V. 6. - 10. - p. e256.
281. Smith A.D., Xuan Z. and Zhang M.Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* - 2008. - V. 9. - p. 128.
282. Sasson A. and Michael T.P. Filtering error from SOLiD Output. *Bioinformatics* - 2010. - V. 26. - 6. - p. 849-50.
283. Li H., Ruan J. and Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* - 2008. - V. 18. - 11. - p. 1851-8.
284. Orian A., Abed M., Kenyagin-Karsenti D., et al. DamID: a methylation-based chromatin profiling approach. *Methods Mol Biol* - 2009. - V. 567. - p. 155-69.
285. Liu X., Noll D.M., Lieb J.D., et al. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res* - 2005. - V. 15. - 3. - p. 421-7.

286. 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. - 4. - p. 622-34.
287. Homer N., Merriman B. and Nelson S.F. BFAST: an alignment tool for large scale genome resequencing. *PLoS One* - 2009. - V. 4. - 11. - p. e7767.
288. Rumble S.M., Lacroute P., Dalca A.V., et al. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* - 2009. - V. 5. - 5. - p. e1000386.
289. Abnizova I., Skelly T., Naumenko F., et al. Statistical comparison of methods to estimate the error probability in short-read Illumina sequencing. *J Bioinform Comput Biol* - 2010. - V. 8. - 3. - p. 579-91.
290. Lunter G. and Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* - 2011. - V. 21. - 6. - p. 936-9.
291. Li R., Li Y., Kristiansen K., et al. SOAP: short oligonucleotide alignment program. *Bioinformatics* - 2008. - V. 24. - 5. - p. 713-4.
292. Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* - 2009. - V. 25. - 14. - p. 1754-60.
293. Jiang H. and Wong W.H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* - 2008. - V. 24. - 20. - p. 2395-6.
294. Zhang Y., Liu T., Meyer C.A., et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* - 2008. - V. 9. - 9. - p. R137.
295. Jothi R., Cuddapah S., Barski A., et al. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* - 2008. - V. 36. - 16. - p. 5221-31.
296. Kim N.K., Jayatillake R.V. and Spouge J.L. NEXT-peak: a normal-exponential two-peak model for peak-calling in ChIP-seq data. *BMC Genomics* - 2013. - V. 14. - p. 349.
297. Qin Z.S., Yu J., Shen J., et al. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics* - 2010. - V. 11. - p. 369.
298. Struhl K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* - 2007. - V. 14. - 2. - p. 103-5.
299. Chen Y., Schmidt B. and Maskell D.L. A hybrid short read mapping accelerator. *BMC Bioinformatics* - 2013. - V. 14. - p. 67.
300. Liu C.M., Wong T., Wu E., et al. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics* - 2012. - V. 28. - 6. - p. 878-9.
301. Liu Y., Schmidt B. and Maskell D.L. CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows-Wheeler transform. *Bioinformatics* - 2012. - V. 28. - 14. - p. 1830-7.
302. Liu X., Lee C.K., Granek J.A., et al. Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Res* - 2006. - V. 16. - 12. - p. 1517-28.
303. Miller J.A. and Widom J. Collaborative competition mechanism for gene activation in vivo. *Mol Cell Biol* - 2003. - V. 23. - 5. - p. 1623-32.
304. Zhang Y., Moqtaderi Z., Rattner B.P., et al. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat Struct Mol Biol* - 2009. - V. 16. - 8. - p. 847-52.
305. Mavrich T.N., Ioshikhes I.P., Venters B.J., et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* - 2008. - V. 18. - 7. - p. 1073-83.
306. Raycroft L., Wu H.Y. and Lozano G. Transcriptional activation by wild-type but not transforming mutants of the p53 anti-oncogene. *Science* - 1990. - V. 249. - 4972. - p. 1049-51.

307. Чумаков П.М. Функция гена p53: выбор между жизнью и смертью. *Биохимия* - 2000. - V. 65. - 1. - p. 34-47.
308. Funk W.D., Pak D.T., Karas R.H., et al. A transcriptionally active DNA-binding site for human p53 protein complexes. *Mol Cell Biol* - 1992. - V. 12. - 6. - p. 2866-71.
309. Lane D.P. and Crawford L.V. T antigen is bound to a host protein in SV40-transformed cells. *Nature* - 1979. - V. 278. - 5701. - p. 261-3.
310. Jenkins J.R., Rudge K. and Currie G.A. Cellular immortalization by a cDNA clone encoding the transformation-associated phosphoprotein p53. *Nature* - 1984. - V. 312. - 5995. - p. 651-4.
311. Levine A.J., Momand J. and Finlay C.A. The p53 tumour suppressor gene. *Nature* - 1991. - V. 351. - 6326. - p. 453-6.
312. Liu X., Miller C.W., Koeffler P.H., et al. The p53 activation domain binds the TATA box-binding polypeptide in Holo-TFIID, and a neighboring p53 domain inhibits transcription. *Mol Cell Biol* - 1993. - V. 13. - 6. - p. 3291-300.
313. Blagosklonny M.V., An W.G., Romanova L.Y., et al. p53 inhibits hypoxia-inducible factor-stimulated transcription. *J Biol Chem* - 1998. - V. 273. - 20. - p. 11995-8.
314. Bhat M.K., Yu C., Yap N., et al. Tumor suppressor p53 is a negative regulator in thyroid hormone receptor signaling pathways. *J Biol Chem* - 1997. - V. 272. - 46. - p. 28989-93.
315. Yu C.L., Driggers P., Barrera-Hernandez G., et al. The tumor suppressor p53 is a negative regulator of estrogen receptor signaling pathways. *Biochem Biophys Res Commun* - 1997. - V. 239. - 2. - p. 617-20.
316. Robertson G., Hirst M., Bainbridge M., et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* - 2007. - V. 4. - 8. - p. 651-7.
317. Leonard W.J. and O'Shea J.J. Jaks and STATs: biological implications. *Annu Rev Immunol* - 1998. - V. 16. - p. 293-322.
318. Ehret G.B., Reichenbach P., Schindler U., et al. DNA binding specificity of different STAT proteins. Comparison of in vitro specificity with natural target sites. *J Biol Chem* - 2001. - V. 276. - 9. - p. 6675-88.
319. Carroll J.S., Liu X.S., Brodsky A.S., et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* - 2005. - V. 122. - 1. - p. 33-43.
320. Carroll J.S. and Brown M. Estrogen receptor target gene: an evolving concept. *Mol Endocrinol* - 2006. - V. 20. - 8. - p. 1707-14.
321. Kang K., Robinson G.W. and Hennighausen L. Comprehensive meta-analysis of Signal Transducers and Activators of Transcription (STAT) genomic binding patterns discerns cell-specific cis-regulatory modules. *BMC Genomics* - 2013. - V. 14. - p. 4.
322. Chan C.S. and Song J.S. CCCTC-binding factor confines the distal action of estrogen receptor. *Cancer Res* - 2008. - V. 68. - 21. - p. 9041-9.
323. Grandori C., Cowley S.M., James L.P., et al. The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu Rev Cell Dev Biol* - 2000. - V. 16. - p. 653-99.
324. Watson J.D., Oster S.K., Shago M., et al. Identifying genes regulated in a Myc-dependent manner. *J Biol Chem* - 2002. - V. 277. - 40. - p. 36921-30.
325. Blackwood E.M. and Eisenman R.N. Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. *Science* - 1991. - V. 251. - 4998. - p. 1211-7.
326. Claassen G.F. and Hann S.R. Myc-mediated transformation: the repression connection. *Oncogene* - 1999. - V. 18. - 19. - p. 2925-33.

327. McMahon S.B., Van Buskirk H.A., Dugan K.A., et al. The novel ATM-related protein TRRAP is an essential cofactor for the c-Myc and E2F oncoproteins. *Cell* - 1998. - V. 94. - 3. - p. 363-74.
328. Amati B., Brooks M.W., Levy N., et al. Oncogenic activity of the c-Myc protein requires dimerization with Max. *Cell* - 1993. - V. 72. - 2. - p. 233-45.
329. Cole M.D. and McMahon S.B. The Myc oncoprotein: a critical evaluation of transactivation and target gene regulation. *Oncogene* - 1999. - V. 18. - 19. - p. 2916-24.
330. Zeller K.I., Jegga A.G., Aronow B.J., et al. An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome Biol* - 2003. - V. 4. - 10. - p. R69.
331. Dang C.V. MYC on the path to cancer. *Cell* - 2012. - V. 149. - 1. - p. 22-35.
332. Menssen A. and Hermeking H. Characterization of the c-MYC-regulated transcriptome by SAGE: identification and analysis of c-MYC target genes. *Proc Natl Acad Sci U S A* - 2002. - V. 99. - 9. - p. 6274-9.
333. Coller H.A., Grandori C., Tamayo P., et al. Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc Natl Acad Sci U S A* - 2000. - V. 97. - 7. - p. 3260-5.
334. Li Z., Van Calcar S., Qu C., et al. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc Natl Acad Sci U S A* - 2003. - V. 100. - 14. - p. 8164-9.
335. Mao D.Y., Watson J.D., Yan P.S., et al. Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. *Curr Biol* - 2003. - V. 13. - 10. - p. 882-6.
336. Frank S.R., Schroeder M., Fernandez P., et al. Binding of c-Myc to chromatin mediates mitogen-induced acetylation of histone H4 and gene activation. *Genes Dev* - 2001. - V. 15. - 16. - p. 2069-82.
337. Chen M., Cui Y.K., Huang W.H., et al. Phosphorylation of estrogen receptor alpha at serine 118 is correlated with breast cancer resistance to tamoxifen. *Oncol Lett* - 2013. - V. 6. - 1. - p. 118-124.
338. Kumar R. and Thompson E.B. The structure of the nuclear hormone receptors. *Steroids* - 1999. - V. 64. - 5. - p. 310-9.
339. Hurtado A., Holmes K.A., Geistlinger T.R., et al. Regulation of ERBB2 by oestrogen receptor-PAX2 determines response to tamoxifen. *Nature* - 2008. - V. 456. - 7222. - p. 663-6.
340. Brzozowski A.M., Pike A.C., Dauter Z., et al. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* - 1997. - V. 389. - 6652. - p. 753-8.
341. Roman-Blas J.A., Castaneda S., Largo R., et al. Osteoarthritis associated with estrogen deficiency. *Arthritis Res Ther* - 2009. - V. 11. - 5. - p. 241.
342. Tsai C.L., Chang L.Y., Chow K.C., et al. Catalase prevents estradiol-induced chondrocyte cytotoxicity. *Life Sci* - 1998. - V. 62. - 13. - p. 1147-52.
343. Pujol P., Rey J.M., Nirde P., et al. Differential expression of estrogen receptor-alpha and -beta messenger RNAs as a potential marker of ovarian carcinogenesis. *Cancer Res* - 1998. - V. 58. - 23. - p. 5367-73.
344. Safe S. and Kim K. Non-classical genomic estrogen receptor (ER)/specificity protein and ER/activating protein-1 signaling pathways. *J Mol Endocrinol* - 2008. - V. 41. - 5. - p. 263-75.
345. Cikes M. Expression of hormone receptors in cancer cells: a hypothesis. *Eur J Cancer* - 1978. - V. 14. - 3. - p. 211-5.
346. Lin C.Y., Vega V.B., Thomsen J.S., et al. Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genet* - 2007. - V. 3. - 6. - p. e87.

347. Welboren W.J., van Driel M.A., Janssen-Megens E.M., et al. ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands. *Embo J* - 2009. - V. 28. - 10. - p. 1418-28.
348. Pan Y.F., Wansa K.D., Liu M.H., et al. Regulation of estrogen receptor-mediated long range transcription via evolutionarily conserved distal response elements. *J Biol Chem* - 2008. - V. 283. - 47. - p. 32977-88.
349. Dean M., Fojo T. and Bates S. Tumour stem cells and drug resistance. *Nat Rev Cancer* - 2005. - V. 5. - 4. - p. 275-84.
350. Weinberg J.M. Topical therapy for actinic keratoses: current and evolving therapies. *Rev Recent Clin Trials* - 2006. - V. 1. - 1. - p. 53-60.
351. Hussain S.P. and Harris C.C. Molecular epidemiology of human cancer: contribution of mutation spectra studies of tumor suppressor genes. *Cancer Res* - 1998. - V. 58. - 18. - p. 4023-37.
352. Cipriano R., Miskimen K.L., Bryson B.L., et al. FAM83B-mediated activation of PI3K/AKT and MAPK signaling cooperates to promote epithelial cell transformation and resistance to targeted therapies. *Oncotarget* - 2013. - V. 4. - 5. - p. 729-38.
353. Saha A., Wittmeyer J. and Cairns B.R. Mechanisms for nucleosome movement by ATP-dependent chromatin remodeling complexes. *Results Probl Cell Differ* - 2006. - V. 41. - p. 127-48.
354. Cairns B.R. Chromatin remodeling: insights and intrigue from single-molecule studies. *Nat Struct Mol Biol* - 2007. - V. 14. - 11. - p. 989-96.
355. Suzuki M.M., Kerr A.R., De Sousa D., et al. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res* - 2007. - V. 17. - 5. - p. 625-31.
356. Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* - 2002. - V. 16. - 1. - p. 6-21.
357. Gribnau J., Hochedlinger K., Hata K., et al. Asynchronous replication timing of imprinted loci is independent of DNA methylation, but consistent with differential subnuclear localization. *Genes Dev* - 2003. - V. 17. - 6. - p. 759-73.
358. Keshet I., Schlesinger Y., Farkash S., et al. Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat Genet* - 2006. - V. 38. - 2. - p. 149-53.
359. Jones P.A. and Baylin S.B. The epigenomics of cancer. *Cell* - 2007. - V. 128. - 4. - p. 683-92.
360. Hoffmann M.J. and Schulz W.A. Causes and consequences of DNA hypomethylation in human cancer. *Biochem Cell Biol* - 2005. - V. 83. - 3. - p. 296-321.
361. Ehrlich M. Cancer-linked DNA hypomethylation and its relationship to hypermethylation. *Curr Top Microbiol Immunol* - 2006. - V. 310. - p. 251-74.
362. Cadieux B., Ching T.T., VandenBerg S.R., et al. Genome-wide hypomethylation in human glioblastomas associated with specific copy number alteration, methylenetetrahydrofolate reductase allele status, and increased proliferation. *Cancer Res* - 2006. - V. 66. - 17. - p. 8469-76.
363. 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. - 17. - p. 8462-9468.
364. Kouzarides T. SnapShot: Histone-modifying enzymes. *Cell* - 2007. - V. 131. - 4. - p. 822.
365. Moss T.J. and Wallrath L.L. Connections between epigenetic gene silencing and human disease. *Mutat Res* - 2007. - V. 618. - 1-2. - p. 163-74.
366. Honda S., Lewis Z.A., Shimada K., et al. Heterochromatin protein 1 forms distinct complexes to direct histone deacetylation and DNA methylation. *Nat Struct Mol Biol* - 2012. - V. 19. - 5. - p. 471-7, S1.

367. Tamaru H. and Selker E.U. A histone H3 methyltransferase controls DNA methylation in Neurospora crassa. *Nature* - 2001. - V. 414. - 6861. - p. 277-83.
368. Merika M., Williams A.J., Chen G., et al. Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription. *Mol Cell* - 1998. - V. 1. - 2. - p. 277-87.
369. Heintzman N.D., Stuart R.K., Hon G., et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* - 2007. - V. 39. - 3. - p. 311-8.
370. Birney E., Stamatoyannopoulos J.A., Dutta A., et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* - 2007. - V. 447. - 7146. - p. 799-816.
371. Bedford D.C., Kasper L.H., Fukuyama T., et al. Target gene context influences the transcriptional requirement for the KAT3 family of CBP and p300 histone acetyltransferases. *Epigenetics* - 2010. - V. 5. - 1. - p. 9-15.
372. Pasini D., Bracken A.P., Jensen M.R., et al. Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. *Embo J* - 2004. - V. 23. - 20. - p. 4061-71.
373. Strahl B.D., Ohba R., Cook R.G., et al. Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in Tetrahymena. *Proc Natl Acad Sci U S A* - 1999. - V. 96. - 26. - p. 14967-72.
374. Kirschmann D.A., Lininger R.A., Gardner L.M., et al. Down-regulation of HP1Hsalpha expression is associated with the metastatic phenotype in breast cancer. *Cancer Res* - 2000. - V. 60. - 13. - p. 3359-63.
375. Norwood L.E., Moss T.J., Margaryan N.V., et al. A requirement for dimerization of HP1Hsalpha in suppression of breast cancer invasion. *J Biol Chem* - 2006. - V. 281. - 27. - p. 18668-76.
376. Tryndyak V.P., Kovalchuk O. and Pogribny I.P. Loss of DNA methylation and histone H4 lysine 20 trimethylation in human breast cancer cells is associated with aberrant expression of DNA methyltransferase 1, Suv4-20h2 histone methyltransferase and methyl-binding proteins. *Cancer Biol Ther* - 2006. - V. 5. - 1. - p. 65-70.
377. Leroy G., Dimaggio P.A., Chan E.Y., et al. A quantitative atlas of histone modification signatures from human cancer cells. *Epigenetics Chromatin* - 2013. - V. 6. - 1. - p. 20.
378. Bannister A.J. and Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res* - 2011. - V. 21. - 3. - p. 381-95.
379. Fraga M.F., Ballestar E., Villar-Garea A., et al. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nat Genet* - 2005. - V. 37. - 4. - p. 391-400.
380. Ting A.H., McGarvey K.M. and Baylin S.B. The cancer epigenome--components and functional correlates. *Genes Dev* - 2006. - V. 20. - 23. - p. 3215-31.
381. Widschwendter M., Fiegl H., Egle D., et al. Epigenetic stem cell signature in cancer. *Nat Genet* - 2007. - V. 39. - 2. - p. 157-8.
382. Lund A.H. and van Lohuizen M. Polycomb complexes and silencing mechanisms. *Curr Opin Cell Biol* - 2004. - V. 16. - 3. - p. 239-46.
383. Feinberg A.P., Ohlsson R. and Henikoff S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet* - 2006. - V. 7. - 1. - p. 21-33.
384. Рязанский С.С. Гвоздев В.А. Короткие РНК и канцерогенез. *Биохимия* - 2008. - V. 73. - 5. - p. 640-655.
385. Kazazian H.H. Jr. and Goodier J.L. LINE drive. retrotransposition and genome instability. *Cell* - 2002. - V. 110. - 3. - p. 277-80.

386. Garcia-Perez J.L., Morell M., Scheys J.O., et al. Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. *Nature* - 2010. - V. 466. - 7307. - p. 769-73.
387. Sinibaldi-Vallebona P., Lavia P., Garaci E., et al. A role for endogenous reverse transcriptase in tumorigenesis and as a target in differentiating cancer therapy. *Genes Chromosomes Cancer* - 2006. - V. 45. - 1. - p. 1-10.
388. Gasior S.L., Wakeman T.P., Xu B., et al. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J Mol Biol* - 2006. - V. 357. - 5. - p. 1383-93.
389. Belancio V.P., Roy-Engel A.M., Pochampally R.R., et al. Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Res* - 2010. - V. 38. - 12. - p. 3909-22.
390. O'Donnell K.A. and Burns K.H. Mobilizing diversity: transposable element insertions in genetic variation and disease. *Mob DNA* - 2010. - V. 1. - 1. - p. 21.
391. Iramaneerat K., Rattanatunyong P., Khemapech N., et al. HERV-K hypomethylation in ovarian clear cell carcinoma is associated with a poor prognosis and platinum resistance. *Int J Gynecol Cancer* - 2011. - V. 21. - 1. - p. 51-7.
392. Faulkner G.J., Kimura Y., Daub C.O., et al. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* - 2009. - V. 41. - 5. - p. 563-71.
393. Hagan C.R. and Rudin C.M. Mobile genetic element activation and genotoxic cancer therapy: potential clinical implications. *Am J Pharmacogenomics* - 2002. - V. 2. - 1. - p. 25-35.
394. Ponicsan S.L., Kugel J.F. and Goodrich J.A. Genomic gems: SINE RNAs regulate mRNA production. *Curr Opin Genet Dev* - 2010. - V. 20. - 2. - p. 149-55.
395. Largaespada D.A. and Collier L.S. Transposon-mediated mutagenesis in somatic cells: identification of transposon-genomic DNA junctions. *Methods Mol Biol* - 2008. - V. 435. - p. 95-108.
396. Sciamanna I., Landriscina M., Pittoggi C., et al. Inhibition of endogenous reverse transcriptase antagonizes human tumor growth. *Oncogene* - 2005. - V. 24. - 24. - p. 3923-31.
397. van Nimwegen M.J. and van de Water B. Focal adhesion kinase: a potential target in cancer therapy. *Biochem Pharmacol* - 2007. - V. 73. - 5. - p. 597-609.
398. Tilghman R.W. and Parsons J.T. Focal adhesion kinase as a regulator of cell tension in the progression of cancer. *Semin Cancer Biol* - 2008. - V. 18. - 1. - p. 45-52.
399. Golubovskaya V.M. Focal adhesion kinase as a cancer therapy target. *Anticancer Agents Med Chem* - 2010. - V. 10. - 10. - p. 735-41.
400. Roberts W.G., Ung E., Whalen P., et al. Antitumor activity and pharmacology of a selective focal adhesion kinase inhibitor, PF-562,271. *Cancer Res* - 2008. - V. 68. - 6. - p. 1935-44.
401. Tanjoni I., Walsh C., Uryu S., et al. PND-1186 FAK inhibitor selectively promotes tumor cell apoptosis in three-dimensional environments. *Cancer Biol Ther* - 2010. - V. 9. - 10. - p. 764-77.
402. Yeh J.E., Toniolo P.A. and Frank D.A. Targeting transcription factors: promising new strategies for cancer therapy. *Curr Opin Oncol* - 2013. - V. 25. - 6. - p. 652-8.
403. Cutroneo K.R. and Ehrlich H. Silencing or knocking out eukaryotic gene expression by oligodeoxynucleotide decoys. *Crit Rev Eukaryot Gene Expr* - 2006. - V. 16. - 1. - p. 23-30.
404. Lin Y.C., Tsai P.H., Lin C.Y., et al. Impact of flavonoids on matrix metalloproteinase secretion and invadopodia formation in highly invasive A431-III cancer cells. *PLoS One* - 2013. - V. 8. - 8. - p. e71903.
405. Moschos S.J., Drogowski L.M., Reppert S.L., et al. Integrins and cancer. *Oncology (Williston Park)* - 2007. - V. 21. - 9 Suppl 3. - p. 13-20.

406. Takahashi K. and Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* - 2006. - V. 126. - 4. - p. 663-76.
407. Kim J., Chu J., Shen X., et al. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* - 2008. - V. 132. - 6. - p. 1049-61.
408. Шутова М.В., *Эпигенетическая характеристика индуцированных плюрипотентных стволовых клеток человека*, in *Лаборатория генетических основ клеточных технологий*. 2011, ИОГен РАН: Москва. p. 88.
409. Gurdon J.B. and Wilmut I. Nuclear transfer to eggs and oocytes. *Cold Spring Harb Perspect Biol* - 2011. - V. 3. - 6. -
410. Pralong D., Trounson A.O. and Verma P.J. Cell fusion for reprogramming pluripotency: toward elimination of the pluripotent genome. *Stem Cell Rev* - 2006. - V. 2. - 4. - p. 331-40.
411. Serov O.L., Matveeva N.M. and Khabarova A.A. Reprogramming mediated by cell fusion technology. *Int Rev Cell Mol Biol* - 2011. - V. 291. - p. 155-90.
412. Takahashi K., Tanabe K., Ohnuki M., et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* - 2007. - V. 131. - 5. - p. 861-72.
413. Waddington C.H., *The Strategy of the Genes; a Discussion of Some Aspects of Theoretical Biology*. 1957, London: Allen and Unwin Ltd. 262
414. Yamanaka S. Elite and stochastic models for induced pluripotent stem cell generation. *Nature* - 2009. - V. 460. - 7251. - p. 49-52.
415. Masui S., Nakatake Y., Toyooka Y., et al. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nat Cell Biol* - 2007. - V. 9. - 6. - p. 625-35.
416. Jaenisch R. and Young R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell* - 2008. - V. 132. - 4. - p. 567-82.
417. Smith A.G. Embryo-derived stem cells: of mice and men. *Annu Rev Cell Dev Biol* - 2001. - V. 17. - p. 435-62.
418. Thomas K.R. and Capecchi M.R. Introduction of homologous DNA sequences into mammalian cells induces mutations in the cognate gene. *Nature* - 1986. - V. 324. - 6092. - p. 34-8.
419. Niwa H., Burdon T., Chambers I., et al. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev* - 1998. - V. 12. - 13. - p. 2048-60.
420. Ying Q.L., Nichols J., Chambers I., et al. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* - 2003. - V. 115. - 3. - p. 281-92.
421. Huangfu D., Maehr R., Guo W., et al. Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nat Biotechnol* - 2008. - V. 26. - 7. - p. 795-7.
422. Shi Y., Desponts C., Do J.T., et al. Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds. *Cell Stem Cell* - 2008. - V. 3. - 5. - p. 568-74.
423. Silva J., Barrandon O., Nichols J., et al. Promotion of reprogramming to ground state pluripotency by signal inhibition. *PLoS Biol* - 2008. - V. 6. - 10. - p. e253.
424. Feng B., Jiang J., Kraus P., et al. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat Cell Biol* - 2009. - V. 11. - 2. - p. 197-203.
425. Chin M.H., Pellegrini M., Plath K., et al. Molecular analyses of human induced pluripotent stem cells and embryonic stem cells. *Cell Stem Cell* - 2010. - V. 7. - 2. - p. 263-9.

426. Boiani M. and Scholer H.R. Regulatory networks in embryo-derived pluripotent stem cells. *Nat Rev Mol Cell Biol* - 2005. - V. 6. - 11. - p. 872-84.
427. Nichols J., Zevnik B., Anastassiadis K., et al. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* - 1998. - V. 95. - 3. - p. 379-91.
428. Boyer L.A., Lee T.I., Cole M.F., et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* - 2005. - V. 122. - 6. - p. 947-56.
429. Loh Y.H., Wu Q., Chew J.L., et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* - 2006. - V. 38. - 4. - p. 431-40.
430. Maherali N., Sridharan R., Xie W., et al. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* - 2007. - V. 1. - 1. - p. 55-70.
431. Okita K., Ichisaka T. and Yamanaka S. Generation of germline-competent induced pluripotent stem cells. *Nature* - 2007. - V. 448. - 7151. - p. 313-7.
432. Wernig M., Meissner A., Foreman R., et al. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* - 2007. - V. 448. - 7151. - p. 318-24.
433. Chambers I., Colby D., Robertson M., et al. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* - 2003. - V. 113. - 5. - p. 643-55.
434. Mitsui K., Tokuzawa Y., Itoh H., et al. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* - 2003. - V. 113. - 5. - p. 631-42.
435. Kaczynski J., Cook T. and Urrutia R. Sp1- and Kruppel-like transcription factors. *Genome Biol* - 2003. - V. 4. - 2. - p. 206.
436. Jiang J., Chan Y.S., Loh Y.H., et al. A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol* - 2008. - V. 10. - 3. - p. 353-60.
437. Ema M., Mori D., Niwa H., et al. Kruppel-like factor 5 is essential for blastocyst development and the normal self-renewal of mouse ESCs. *Cell Stem Cell* - 2008. - V. 3. - 5. - p. 555-67.
438. Hall J., Guo G., Wray J., et al. Oct4 and LIF/Stat3 additively induce Kruppel factors to sustain embryonic stem cell self-renewal. *Cell Stem Cell* - 2009. - V. 5. - 6. - p. 597-609.
439. Nakagawa M., Koyanagi M., Tanabe K., et al. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat Biotechnol* - 2008. - V. 26. - 1. - p. 101-6.
440. Ivanova N., Dobrin R., Lu R., et al. Dissecting self-renewal in stem cells with RNA interference. *Nature* - 2006. - V. 442. - 7102. - p. 533-8.
441. Galan-Caridad J.M., Harel S., Arenzana T.L., et al. Zfx controls the self-renewal of embryonic and hematopoietic stem cells. *Cell* - 2007. - V. 129. - 2. - p. 345-57.
442. Wang J., Rao S., Chu J., et al. A protein interaction network for pluripotency of embryonic stem cells. *Nature* - 2006. - V. 444. - 7117. - p. 364-8.
443. Sladek R., Bader J.A. and Giguere V. The orphan nuclear receptor estrogen-related receptor alpha is a transcriptional regulator of the human medium-chain acyl coenzyme A dehydrogenase gene. *Mol Cell Biol* - 1997. - V. 17. - 9. - p. 5400-9.
444. Bieda M., Xu X., Singer M.A., et al. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res* - 2006. - V. 16. - 5. - p. 595-605.
445. Cartwright P., McLean C., Sheppard A., et al. LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. *Development* - 2005. - V. 132. - 5. - p. 885-96.

446. Watt F.M. and Huck W.T. Role of the extracellular matrix in regulating stem cell fate. *Nat Rev Mol Cell Biol* - 2013. - V. 14. - 8. - p. 467-73.
447. Kim J.B., Greber B., Arauzo-Bravo M.J., et al. Direct reprogramming of human neural stem cells by OCT4. *Nature* - 2009. - V. 461. - 7264. - p. 649-3.
448. Davis R.L., Weintraub H. and Lassar A.B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* - 1987. - V. 51. - 6. - p. 987-1000.
449. Gu P., Goodwin B., Chung A.C., et al. Orphan nuclear receptor LRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development. *Mol Cell Biol* - 2005. - V. 25. - 9. - p. 3492-505.
450. Zhou Q., Chipperfield H., Melton D.A., et al. A gene regulatory network in mouse embryonic stem cells. *Proc Natl Acad Sci U S A* - 2007. - V. 104. - 42. - p. 16438-43.
451. Tay Y.M., Tam W.L., Ang Y.S., et al. MicroRNA-134 modulates the differentiation of mouse embryonic stem cells, where it causes post-transcriptional attenuation of Nanog and LRH1. *Stem Cells* - 2008. - V. 26. - 1. - p. 17-29.
452. Zhao T., Zhang Z.N., Rong Z., et al. Immunogenicity of induced pluripotent stem cells. *Nature* - 2011. - V. 474. - 7350. - p. 212-5.
453. Tang C., Weissman I.L. and Drukker M. The safety of embryonic stem cell therapy relies on teratoma removal. *Oncotarget* - 2012. - V. 3. - 1. - p. 7-8.
454. Kaneko S. and Yamanaka S. To be immunogenic, or not to be: that's the iPSC question. *Cell Stem Cell* - 2013. - V. 12. - 4. - p. 385-6.
455. Araki R., Uda M., Hoki Y., et al. Negligible immunogenicity of terminally differentiated cells derived from induced pluripotent or embryonic stem cells. *Nature* - 2013. - V. 494. - 7435. - p. 100-4.
456. Lister R., Pelizzola M., Kida Y.S., et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* - 2011. - V. 471. - 7336. - p. 68-73.
457. Lindgren A.G., Natsuhara K., Tian E., et al. Loss of Pten causes tumor initiation following differentiation of murine pluripotent stem cells due to failed repression of Nanog. *PLoS One* - 2011. - V. 6. - 1. - p. e16478.
458. Vierbuchen T., Ostermeier A., Pang Z.P., et al. Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* - 2010. - V. 463. - 7284. - p. 1035-41.
459. Zhou Q., Brown J., Kanarek A., et al. In vivo reprogramming of adult pancreatic exocrine cells to beta-cells. *Nature* - 2008. - V. 455. - 7213. - p. 627-32.
460. Ieda M., Fu J.D., Delgado-Olguin P., et al. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* - 2010. - V. 142. - 3. - p. 375-86.
461. Pang Z.P., Yang N., Vierbuchen T., et al. Induction of human neuronal cells by defined transcription factors. *Nature* - 2011. - V. 476. - 7359. - p. 220-3.
462. Nethercott H.E., Brick D.J. and Schwartz P.H. Derivation of induced pluripotent stem cells by lentiviral transduction. *Methods Mol Biol* - 2011. - V. 767. - p. 67-85.
463. Okita K., Yamakawa T., Matsumura Y., et al. An efficient nonviral method to generate integration-free human-induced pluripotent stem cells from cord blood and peripheral blood cells. *Stem Cells* - 2013. - V. 31. - 3. - p. 458-66.
464. Davis R.P., Nemes C., Varga E., et al. Generation of induced pluripotent stem cells from human foetal fibroblasts using the Sleeping Beauty transposon gene delivery system. *Differentiation* - 2013. - V. 86. - 1-2. - p. 30-7.
465. Yusa K., Rad R., Takeda J., et al. Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. *Nat Methods* - 2009. - V. 6. - 5. - p. 363-9.
466. Hayes M. and Zavazava N. Strategies to generate induced pluripotent stem cells. *Methods Mol Biol* - 2013. - V. 1029. - p. 77-92.
467. Ladewig J., Mertens J., Kesavan J., et al. Small molecules enable highly efficient neuronal conversion of human fibroblasts. *Nat Methods* - 2012. - V. 9. - 6. - p. 575-8.

468. Hou P., Li Y., Zhang X., et al. Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. *Science* - 2013. - V. 341. - 6146. - p. 651-4.
469. Chen J., Wang G., Lu C., et al. Synergetic cooperation of microRNAs with transcription factors in iPS cell generation. *PLoS One* - 2012. - V. 7. - 7. - p. e40849.
470. Chin M.H., Mason M.J., Xie W., et al. Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* - 2009. - V. 5. - 1. - p. 111-23.
471. Meissner A., Mikkelsen T.S., Gu H., et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* - 2008. - V. 454. - 7205. - p. 766-70.
472. de Wit E., Bouwman B.A., Zhu Y., et al. The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* - 2013. - V. 501. - 7466. - p. 227-31.
473. Nora E.P., Dekker J. and Heard E. Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? *Bioessays* - 2013. - V. 35. - 9. - p. 818-28.
474. Cremer C., Cremer T. and Gray J.W. Induction of chromosome damage by ultraviolet light and caffeine: correlation of cytogenetic evaluation and flow karyotype. *Cytometry* - 1982. - V. 2. - 5. - p. 287-90.
475. Панова А.В. Некрасов Е.Д., Лагарькова М.А., Киселёв С.Л., Богомазова А.Н. Поздняя репликация инактивированной х-хромосомы в плюрипотентных стволовых клетках человека не зависит от степени компактизации ее хромосомной территории. *Acta naturae* - 2013. - V. 5. - 2(17). - p. 55-63.
476. Shopland L.S., Lynch C.R., Peterson K.A., et al. Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. *J Cell Biol* - 2006. - V. 174. - 1. - p. 27-38.
477. Simonis M., Klous P., Splinter E., et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* - 2006. - V. 38. - 11. - p. 1348-54.
478. Solovei I., Kreysing M., Lanctot C., et al. Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. *Cell* - 2009. - V. 137. - 2. - p. 356-68.
479. Dekker J., Marti-Renom M.A. and Mirny L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* - 2013. - V. 14. - 6. - p. 390-403.
480. Gilbert N., Boyle S., Fiegler H., et al. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* - 2004. - V. 118. - 5. - p. 555-66.
481. Morey C., Da Silva N.R., Perry P., et al. Nuclear reorganisation and chromatin decondensation are conserved, but distinct, mechanisms linked to Hox gene activation. *Development* - 2007. - V. 134. - 5. - p. 909-19.
482. 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. - 5950. - p. 289-93.
483. Sanyal A., Bau D., Marti-Renom M.A., et al. Chromatin globules: a common motif of higher order chromosome structure? *Curr Opin Cell Biol* - 2011. - V. 23. - 3. - p. 325-31.
484. Dekker J., Rippe K., Dekker M., et al. Capturing chromosome conformation. *Science* - 2002. - V. 295. - 5558. - p. 1306-11.
485. Orlando V., Strutt H. and Paro R. Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods* - 1997. - V. 11. - 2. - p. 205-14.
486. Jackson V. Formaldehyde cross-linking for studying nucleosomal dynamics. *Methods* - 1999. - V. 17. - 2. - p. 125-39.

487. Fujita N. and Wade P.A. Use of bifunctional cross-linking reagents in mapping genomic distribution of chromatin remodeling complexes. *Methods* - 2004. - V. 33. - 1. - p. 81-5.
488. Wurtele H. and Chartrand P. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Res* - 2006. - V. 14. - 5. - p. 477-95.
489. Zhao Z., Tavoosidana G., Sjolinder M., et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* - 2006. - V. 38. - 11. - p. 1341-7.
490. Dostie J., Richmond T.A., Arnaout R.A., et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* - 2006. - V. 16. - 10. - p. 1299-309.
491. Gavrilov A., Eivazova E., Priozhkova I., et al. Chromosome conformation capture (from 3C to 5C) and its ChIP-based modification. *Methods Mol Biol* - 2009. - V. 567. - p. 171-88.
492. Fullwood M.J., Han Y., Wei C.L., et al. Chromatin interaction analysis using paired-end tag sequencing. *Curr Protoc Mol Biol* - 2010. - V. Chapter 21. - p. Unit 21 15 1-25.
493. de Wit E. and de Laat W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* - 2012. - V. 26. - 1. - p. 11-24.
494. Nagano T., Lubling Y., Stevens T.J., et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* - 2013. - V. 502. - 7469. - p. 59-64.
495. Shaw P.J. Mapping chromatin conformation. *F1000 Biol Rep* - 2010. - V. 2. -
496. Beatty B. M.S., Squire J., *Oxford University Press*. Medical 2002, Oxford: Oxford University Press. 255.
497. Bian Q. and Belmont A.S. Revisiting higher-order and large-scale chromatin organization. *Curr Opin Cell Biol* - 2012. - V. 24. - 3. - p. 359-66.
498. Ryba T., Hiratani I., Lu J., et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* - 2010. - V. 20. - 6. - p. 761-70.
499. Mirny L.A. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res* - 2011. - V. 19. - 1. - p. 37-51.
500. Jacob F., Perrin D., Sanchez C., et al. [Operon: a group of genes with the expression coordinated by an operator]. *C R Hebd Seances Acad Sci* - 1960. - V. 250. - p. 1727-9.
501. Pauli D., Tonka C.H. and Ayme-Southgate A. An unusual split Drosophila heat shock gene expressed during embryogenesis, pupation and in testis. *J Mol Biol* - 1988. - V. 200. - 1. - p. 47-53.
502. Zorio D.A., Cheng N.N., Blumenthal T., et al. Operons as a common form of chromosomal organization in C. elegans. *Nature* - 1994. - V. 372. - 6503. - p. 270-2.
503. Cook P.R. The organization of replication and transcription. *Science* - 1999. - V. 284. - 5421. - p. 1790-5.
504. Cremer T. and Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* - 2001. - V. 2. - 4. - p. 292-301.
505. van Steensel B. and Dekker J. Genomics tools for unraveling chromosome architecture. *Nat Biotechnol* - 2010. - V. 28. - 10. - p. 1089-1095.
506. Cope N.F., Fraser P. and Eskiw C.H. The yin and yang of chromatin spatial organization. *Genome Biol* - 2010. - V. 11. - 3. - p. 204.
507. Handoko L., Xu H., Li G., et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* - 2011. - V. 43. - 7. - p. 630-8.
508. Bau D., Sanyal A., Lajoie B.R., et al. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* - 2011. - V. 18. - 1. - p. 107-14.

509. Miller L.D., Smeds J., George J., et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A* - 2005. - V. 102. - 38. - p. 13550-13555.
510. Ishikawa K., Nishihara H., Ozawa S., et al. Focal macular electroretinograms after photodynamic therapy combined with intravitreal bevacizumab. *Graefes Arch Clin Exp Ophthalmol* - 2010. - V. 249. - 2. - p. 273-280.
511. Kalathur R.K., Gagniere N., Berthommier G., et al. RETINOBASE: a web database, data mining and analysis platform for gene expression data on retina. *BMC Genomics* - 2008. - V. 9. - p. 208.
512. Ivanisenko V.A., Pintus S.S., Grigorovich D.A., et al. PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res* - 2005. - V. 33. - Database issue. - D183-7.
513. Shadeo A. and Lam W.L. Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Res* - 2006. - V. 8. - 1. - R9.
514. Xu H., Handoko L., Wei X., et al. A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics* - 2010. - V. 26. - 9. - p. 1199-1204.
515. Satchwell S.C., Drew H.R. and Travers A.A. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* - 1986. - V. 191. - 4. - p. 659-675.
516. Clarke N.D. and Granek J.A. Rank order metrics for quantifying the association of sequence features with gene regulation. *Bioinformatics* - 2003. - V. 19. - 2. - p. 212-218.
517. Hou C., Dale R. and Dean A. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci U S A* - 2010. - V. 107. - 8. - p. 3651-3656.
518. Cai S., Lee C.C. and Kohwi-Shigematsu T. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nat Genet* - 2006. - V. 38. - 11. - p. 1278-1288.
519. Schoenfelder S., Sexton T., Chakalova L., et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* - 2010. - V. 42. - 1. - p. 53-61.
520. Kilpelainen T.O., Zillikens M.C., Stancakova A., et al. Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. *Nat Genet* - 2011. - V. 43. - 8. - p. 753-760.
521. Kasowski M., Grubert F., Heffelfinger C., et al. Variation in transcription factor binding among humans. *Science* - 2010. - V. 328. - 5975. - p. 232-235.

ПРИЛОЖЕНИЕ

Таблица П.1

Международные проекты геномных исследований

Название программы, ссылка	Интернет-адрес, краткое описание
1000 Genomes Project	http://www.1000genomes.org поиск геномных вариаций в 26 популяциях (~ 2500 персональных геномов)
PGP: personal genome project	http://www.personalgenomes.org/ Персональный Геномный Проект. Изучение вклада наследственности и среды в проявление различных признаков; секвенирование геномов
gEUVADIS (Genetic European Variation in Disease)	http://www.geuvadis.org Медицинская интерпретация данных секвенирования RNA-seq и ExonSeq для моно- и полигенных заболеваний
ESGI (European Sequencing and Genotyping Infrastructure)	http://www.esgi-infrastructure.eu организация NGS инфраструктуры для европейского научного сообщества
IHEC (International Human Epigenome Consortium)	www.ihec-epigenomes.org характеристика эпигенома человека, ~1000 образцов
Программы исследования рака, поиск диагностических маркеров, анализ и выбор лечения	
ICGC (International Cancer Genome Consortium)	icgc.org/ Международный консорциум исследования рака
Treat 1000	http://www.treat1000.org разные виды рака
OncoTrack	http://www.oncotrack.eu рак прямой кишки
CAGEKID	http://www.cng.fr/cagekid/ рак почки
TREAT 20	http://cccc.charite.de/ меланома
TCGA (The Cancer Genome Atlas)	http://cancergenome.nih.gov/ каталог генов рака

Таблица П.2

Базы и репозитарии геномных данных и данных экспрессии генов

База данных	URL линк
NCBI (GenBank)	http://www.ncbi.nlm.nih.gov/
EMBL-EBI	http://www.ebi.ac.uk/
DDBJ	http://www.ddbj.nig.ac.jp/
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/
Ensembl	http://www.ensembl.org/
UCSC Genome Browser	http://genome.ucsc.edu/
NCBI MapViewer	http://www.ncbi.nlm.nih.gov/mapview/
ENCODE (project)	http://www.genome.gov/10005107
1000 Genomes	http://www.1000genomes.org/
Rat Genome Database	http://rgd.mcw.edu/
Mouse Genome Informatics	http://www.informatics.jax.org/
ZFIN, Zebrafish Model Organism Database	http://zfin.org
SGD, Saccharomyces Genome Database	http://www.yeastgenome.org/
HapMap	http://www.hapmap.org
GEO, Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/geo/
ArrayExpress	http://www.ebi.ac.uk/microarray-
BioGPS	http://biogps.org

Таблица П.3

15-буквенный нуклеотидный код (Nomenclature Committee-IUB, 1986)

15-буквенный код	Сопоставляемая группа нуклеотидов	Мнемоническое правило	Название или общее свойство
A	A		Аденин
T	T		Тимин
G	G		Гуанин
C	C		Цитозин
W	A,T	WEAK	Слабые взаимодействия
R	A,G	PURINE	Пурины
M	A,C	AMINO- (пол. заряд)	Амино-группа
K	T,G	KETO- (отриц. заряд)	Кето-группа
Y	T,C	PYRIMIDINE	пиримидины
S	G,C	STRONG	Сильные взаимодействия
B	T,G,C	не A	
V	A,G,C	не T	
H	A,T,C	не G	
D	A,T,G	не C	
N	A,T,G,C	Любой	

Таблица П.4

Компании, разрабатывающие программные продукты в области системной биологии, относящихся к секвенированию и иммунопреципитации хроматина

Компании (код страны)	**Основные продукты, бренды**	**На какие отрасли ориентирована продукция**	**Интернет адрес компании**
Accelrys Inc. (US)	Разрабатывает компьютерные программы для системной биологии Discovery Studio	Биоинформатика, фармакология	http://accelrys.com/
Affymetrix, Inc. (US)	Разрабатывает геномные технологии, микрочипы и программы геномного анализа	Геномика, биоинформатика	http://www.affymetrix.com
Applied Biosystems (US)	SOLiD™ Sequencing Разрабатывает компьютерные программы для системной биологии, секвенирования	Биоинформатика, фармакология, медицина	http://www.appliedbiosystems.com/
Ariadne Genomics, Inc. (US)	Pathway Studio® Интерактивная компьютерная система для реконструкции и визуализации взаимодействия между биологическими объектами и процессами.	Биоинформатика, медицина и токсикогеномика	http://www.ariadnegenomics.com/
Artificial Intelligence Lab. (US)	BioPathway® Программные продукты для биоинформатики.	Биоинформатика, медицина.	http://ai.eller.arizona.edu
Beyond Genomics или BG Medicine, Inc. (US)	Molecular Phenotypes™ Платформа для дизайна лекарств	Биоинформатика, фармакология	www.bg-medicine.com/
BGI (Beijing Genome Institute) (CN)	Разрабатывает решения для геномики, включая обработку и хранение данных секвенирования	Геномное секвенирование, биоинформатика	http://www.cloud-sequencing.com/ www.genomics.cn www.bgisequence.com/eu
Biobase (DE)	Разрабатывает компьютерные базы данных в области биотехнологий	Биоинформатика, фармакология, медицина	http://www.biobase-international.com/
Bristol-Myers Squibb (US)	Фармацевтическая компания, разрабатывает компьютерные средства прикладной геномики	Фармацевтика, биоинформатика	http://www.bms.com
Celera (US)	Разрабатывает решения для геномного секвенирования, включая персональную	Геномное секвенирование, биоинформатика	https://www.celera.com/

Компании (код страны)	Основные продукты, бренды	На какие отрасли ориентирована продукция	Интернет адрес компании
	медицину		
CLC bio (DK)	CLC Genomics Workbench Программное обеспечение компьютерной геномики	Геномика	http://www.clcbio.com/
Complete Genomics Inc. (US)	Разрабатывает компьютерные программы для системной биологии	Биоинформатика, фармакология, медицина	http://www.completegenomics.com/
DNASTAR, Inc. (US)	Lasergene Разрабатывает компьютерные программы для системной биологии	Геномика, биоинформатика	http://www.dnastar.com/
Eli Lilly and Company (US)	Фармацевтическая корпорация разрабатывает лекарственные средства, используя интегральные решения геномики	Фармакология, медицина, биоинформатика	http://www.lilly.ru/Nitro/
Gateway Inc. (US)	Разрабатывает компьютерные программы для системной биологии	Биоинформатика, фармакология, медицина	http://www.gateway.com/
Gene Network Sciences (US, UK)	Разрабатывает компьютерные модели на уровне клетки и организма	Биоинформатика, фармакология, системная биология	http://www.gnsbio-tech.com/
Genedata AG (CH, DE, US)	Компьютерные программы для системной биологии и для поиска новых лекарственных средств: Genedata Phylosopher®, Genedata Screener®, Genedata Expressionist®	Биоинформатика, фармакология, медицина	http://www.genedata.com/
GeneGo (US) (Thomson Reuters)	Компьютерные программы для интеграции и компьютерного анализа экспериментальных данных. Продукты: MetaCore™ - анализ микрочиповых данных	Биоинформатика, фармакология, медицина	http://www.genego.com/
Genomatica (US)	Продукты: SimPheny™ Modeling Platform Predictive Metabolic Models, Model Development, Research and Discovery	Биотехнология, Биоинформатика, медицины	http://www.genomatica.com/index.shtml
Helicos BioSciences	Продукты: Helicos Технологии секвенирования и	Геномика	www.helicosbio.com

Компании (код страны)	**Основные продукты, бренды**	**На какие отрасли ориентирована продукция**	**Интернет адрес компании**
Corporation,	геномного анализа		
Human Genome Sciences, Inc.	Разрабатывает программы геномного анализа	Фармацевтика Геномика, биоинформатика	http://www.hgsi.com/
IBM Corp. (US)	Продукты для молекулярного компьютерного дизайна (CAMD), фармакологии (IBM BioPharmaceutical Solution)	Биоинформатика, медицина, биотехнология.	http://www.ibm.com/us/
Illumina, Inc (US)	Illumina Solexa. HiSeq Технологии секвенирования и компьютерные системы анализа геномных данных	Геномика, секвенирование ДНК, генотипирование	http://www.illumina.com/
Incyte Genomics, Inc. (US)	Разрабатывает компьютерные программы для геномики	Фармацевтика, геномика	http://www.incyte.com/
Ion Torrent Systems, Inc.,	Ion Torrent Технологии секвенирования и геномного анализа	Геномика	http://www.iontorrent.com/
Ip Genesis Inc (US)	Разрабатывает компьютерные программы для системной биологии	Биоинформатика, фармакология, медицина	http://www.ipgenesis.com/
Pacific Biosciences (US)	PacBio SMRT (Single Molecule Real Time). Технологии секвенирования	Геномика	http://www.pacificbiosciences.com/
PREMIER Biosoft (US)	AlleleID Программы биоинформатики, дизайн микрочипов	Биоинформатика	http://www.premierbiosoft.com/
QIAGEN N.V. (NL)	Разрабатывает геномные диагностики	Геномная диагностика	http://www.qiagen.com/
Real Time Genomics (US)	RTG Investigator Разрабатывает программы по геномике растений	Геномика растений биоинформатика	http://www.realtimegenomics.com/
Real Time Genomics Inc. (US)	Разрабатывает компьютерные программы для системной биологии	Биоинформатика, фармакология, медицина	http://www.realtimegenomics.com/
Rosetta Genomics, Ltd (IL)	Медицинские диагностики, связанные с микроРНК, компьютерные алгоритмы	Медицинская диагностика, биоинформатика	https://www.rosettagenomics.com
RURO Inc. (US)	Программные средства для фармацевтических и биотехнологических компаний	Биоинформатика, Биотехнология, фармакология	http://www.ruro.com/

Таблица П.5

Компьютерная симуляция числа специфичных кластеров для ChIP-seq библиотеки Nanog используя фиксированный процент специфичных последовательностей

Число прочтений ChIP-seq	Порог высоты пика	Макс. число сайтов N'	Мин. число сайтов N'' (для порога +1)	Аппрокси-мация для N'	Аппрокси-мация для N''
100000	3	2496	524	4770,8	1128,2
200000	3	7534	1432	6693,5	2053,6
300000	3	14477	2624	7732,3	2826,4
400000	3	22801	4033	8382,8	3481,5
500000	3	32708	5622	8828,4	4043,8
600000	4	7368	3493	9152,8	4531,8
700000	4	9508	4306	9399,4	4959,2
800000	4	11767	5190	9593,3	5336,8
900000	4	14280	6141	9749,8	5672,7
1000000	4	16954	7082	9878,6	5973,5
1250000	5	9589	6051	10119,4	6603,7
1500000	5	12438	7702	10286,6	7103,3
1750000	6	9431	6965	10409,4	7509,2
2000000	6	11293	8259	10503,4	7845,3
2250000	7	9509	7518	10577,7	8128,3
2500000	7	10827	8517	10638	8369,9
2750000	8	9615	7833	10687,7	8578,4
3000000	8	10685	8768	10729,6	8760,4
3250000	9	9692	8209	10765,3	8920,4
3500000	9	10588	8981	10796	9062,4
3750000	9	11506	9725	10822,8	9189,1
4000000	10	10564	9086	10846,4	9302,9
4250000	10	11359	9787	10867,3	9405,7
4500000	11	10490	9172	10885,9	9499
4750000	11	11210	9797	10902,6	9584,1
5000000	12	10434	9233	10917,7	9661,9
5250000	12	11065	9819	10931,4	9733,5
5500000	12	11694	10401	10943,8	9799,4
5750000	13	10983	9855	10955,3	9860,5
6000000	13	11544	10360	10965,7	9917,1
6250000	14	10887	9827	10975,4	9969,7
6500000	14	11413	10317	10984,3	10018,8
6684737	15	10688	9769	10990,5	10053

Таблица П.6

Кластеры сайтов связывания транскрипционных факторов в геноме мыши

Хромосомная локализация кластеров ТФ (геном мыши, mm8)	Число ТФ в кластере	Список ТФ в кластере
chr17:35111908-35112293	11	E2f1;Esrrb;Klf4;Nanog;n-Myc;Oct4;Smad1; Sox2; Stat3; Tcfcp2l1; Zfx;
chr13:34117205-34117536	11	c-Myc;E2f1;Esrrb;Klf4;Nanog;n-Myc;Oct4;Smad1;Sox2; Stat3;Tcfcp2l1;
chr8:109824576-109824894	10	c-Myc;E2f1;Esrrb;Klf4;Nanog;n-Myc;Oct4;Stat3; Tcfcp2l1; Zfx;
chr7:109401154-109401525	10	c-Myc;E2f1;Esrrb;Klf4;Nanog;n-Myc;Oct4;Sox2;Stat3 ;Tcfcp2l1;
chr3:103275401-103275736	10	E2f1;Esrrb;Klf4;Nanog;n-Myc;Oct4;Smad1;Sox2;Stat3; Tcfcp2l1;
chr2:51894032-51894327	10	c-Myc;E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3; Tcfcp2l1;
chr2:154119310-154119705	10	E2f1;Esrrb;Klf4;Nanog;n-Myc;Oct4;Smad1;Sox2;Stat3; Tcfcp2l1;
chr12:87390821-87391206	10	c-Myc;Esrrb;Klf4;Nanog;n-Myc;Oct4;Smad1;Sox2;Stat3; Tcfcp2l1;
chr1:135510904-135511224	10	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3; Tcfcp2l1; Zfx;
chr9:7634035-7634268	9	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr8:125409342-125409785	9	c-Myc;E2f1;Klf4;n-Myc;Oct4;Smad1;Stat3;Tcfcp2l1;Zfx;
chr7:80003756-80004207	9	c-Myc;E2f1;Esrrb;Klf4;n-Myc;Oct4;Smad1;Sox2;Stat3;
chr7:107130818-107131119	9	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr5:66003875-66004222	9	c-Myc;CTCF;E2f1;Klf4;Nanog;n-Myc;Stat3;Tcfcp2l1;Zfx;
chr5:28540606-28541046	9	E2f1;Esrrb;Klf4;Nanog;n-Myc;Oct4;Smad1;Sox2;Tcfcp2l1;
chr4:140842092-140842362	9	Esrrb;Klf4;Nanog;n-Myc;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr2:167806261-167806522	9	CTCF;E2f1;Esrrb;Klf4;n-Myc;Sox2;Stat3;Tcfcp2l1;Zfx;
chr2:11520855-11521280	9	c-Myc;CTCF;E2f1;Esrrb;Klf4;n-Myc;Oct4;Tcfcp2l1;Zfx;
chr19:28248708-28249060	9	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr16:30740955-30741288	9	c-Myc;E2f1;Esrrb;Nanog;n-Myc;Oct4;Sox2;Stat3;Tcfcp2l1;
chr15:103348001-103348418	9	E2f1;Esrrb;Klf4;Nanog;n-Myc;Oct4;Sox2;Stat3;Tcfcp2l1;
chr14:7197154-7197488	9	E2f1;Esrrb;Klf4;Nanog;n-Myc;Oct4;Smad1;Sox2;Stat3;
chr13:51858446-51858651	9	c-Myc;E2f1;Esrrb;Nanog;n-Myc;Oct4;Smad1;Sox2; Tcfcp2l1;
chr10:79801772-79802078	9	c-Myc;E2f1;Esrrb;Nanog;n-Myc;Oct4;Smad1;Sox2;Stat3;
chr9:95321318-95321663	8	c-Myc;E2f1;Esrrb;Klf4;n-Myc;Oct4;Stat3;Tcfcp2l1;

Хромосомная локализация кластеров ТФ (геном мыши, mm8)	Число ТФ в кластере	Список ТФ в кластере
chr9:59455113-59455525	8	c-Myc;CTCF;E2f1;Esrrb;Nanog;n-Myc;Tcfcp2l1;Zfx;
chr9:58077455-58077745	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr9:49503743-49504035	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr9:45100253-45100560	8	Esrrb;Nanog;n-Myc;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr9:31845065-31845461	8	E2f1;Esrrb;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr8:98032623-98032944	8	c-Myc;Klf4;n-Myc;Oct4;Smad1;Sox2;Tcfcp2l1;Zfx;
chr8:74314565-74314855	8	c-Myc;CTCF;E2f1;Esrrb;Nanog;n-Myc;Oct4;Zfx;
chr8:72830503-72830797	8	c-Myc;CTCF;E2f1;Klf4;Nanog;n-Myc;Oct4;Stat3;
chr8:60625069-60625245	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Tcfcp2l1;
chr8:38017714-38017951	8	c-Myc;E2f1;Esrrb;Klf4;Nanog;n-Myc;Stat3;Tcfcp2l1;
chr8:35065332-35065678	8	c-Myc;CTCF;E2f1;Klf4;n-Myc;Oct4;Tcfcp2l1;Zfx;
chr8:109340565-109340749	8	Esrrb;Klf4;Nanog;n-Myc;Oct4;Stat3;Tcfcp2l1;Zfx;
chr7:89850791-89851171	8	E2f1;Esrrb;Klf4;Nanog;n-Myc;Sox2;Stat3;Zfx;
chr7:65752427-65752706	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Stat3;Tcfcp2l1;Zfx;
chr7:63836008-63836397	8	c-Myc;E2f1;Klf4;Nanog;Oct4;Sox2;Tcfcp2l1;Zfx;
chr7:11914490-11914861	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr6:100343423-100343853	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Tcfcp2l1;
chr5:65143886-65144141	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Zfx;
chr5:33852254-33852637	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Sox2;Tcfcp2l1;Zfx;
chr5:25064765-25065069	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Tcfcp2l1;
chr5:140584614-140584847	8	c-Myc;CTCF;Esrrb;Nanog;n-Myc;Oct4;Sox2;Tcfcp2l1;
chr5:135227411-135227772	8	E2f1;Esrrb;Nanog;Oct4;Smad1;Sox2;Tcfcp2l1;Zfx;
chr4:55496510-55496769	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr4:141342917-141343149	8	c-Myc;Klf4;Nanog;n-Myc;Smad1;Sox2;Stat3;Tcfcp2l1;
chr4:135418821-135419206	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Tcfcp2l1;
chr4:123126817-123127059	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr4:10859507-10859809	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr4:104721965-104722342	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Sox2;Stat3;Tcfcp2l1;
chr3:8991807-8992082	8	Esrrb;Nanog;n-Myc;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr3:88617262-88617610	8	c-Myc;E2f1;Klf4;n-Myc;Oct4;Sox2;Stat3;Zfx;
chr3:30837684-30837956	8	E2f1;Esrrb;Klf4;Nanog;n-Myc;Smad1;Stat3;Tcfcp2l1;
chr3:18816053-18816445	8	CTCF;Esrrb;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr3:137717558-137717935	8	E2f1;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr3:121569294-121569705	8	E2f1;Esrrb;Klf4;n-Myc;Oct4;Sox2;Tcfcp2l1;Zfx;
chr2:91944869-91945153	8	Esrrb;Nanog;n-Myc;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr2:38490788-38491072	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Sox2;Tcfcp2l1;Zfx;
chr2:180156881-180157248	8	CTCF;Esrrb;Klf4;Nanog;n-Myc;Oct4;Stat3;Tcfcp2l1;
chr2:172268419-172268904	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;

Хромосомная локализация кластеров ТФ (геном мыши, mm8)	Число ТФ в кластере	Список ТФ в кластере
chr2:144406802-144407032	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Tcfcp2l1;
chr2:143962051-143962326	8	c-Myc;E2f1;Esrrb;Klf4;Nanog;n-Myc;Oct4;Zfx;
chr19:23162129-23162390	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr17:84259730-84260014	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Tcfcp2l1;
chr17:83912260-83912513	8	E2f1;Klf4;Nanog;Oct4;Sox2;Stat3;Tcfcp2l1;Zfx;
chr17:35098426-35098706	8	E2f1;Esrrb;Nanog;n-Myc;Oct4;Sox2;Stat3;Tcfcp2l1;
chr16:9046489-9046790	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr16:3992957-3993199	8	E2f1;Esrrb;Nanog;n-Myc;Oct4;Stat3;Tcfcp2l1;Zfx;
chr16:37570744-37570968	8	c-Myc;E2f1;Esrrb;Klf4;Nanog;n-Myc;Stat3;Tcfcp2l1;
chr16:35507677-35508065	8	c-Myc;Esrrb;Klf4;Nanog;n-Myc;Oct4;Sox2;Zfx;
chr16:30561138-30561553	8	CTCF;E2f1;Esrrb;Nanog;Oct4;Smad1;Sox2;Tcfcp2l1;
chr16:13694129-13694348	8	c-Myc;E2f1;Esrrb;Klf4;n-Myc;Oct4;Tcfcp2l1;Zfx;
chr15:98821097-98821353	8	CTCF;Esrrb;Klf4;Nanog;n-Myc;Oct4;Smad1;Stat3;
chr15:88536698-88536984	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr15:61922315-61922659	8	E2f1;Esrrb;Klf4;Nanog;Smad1;Sox2;Stat3;Tcfcp2l1;
chr15:50730530-50730745	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr14:75249804-75250031	8	CTCF;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Tcfcp2l1;
chr14:62456027-62456215	8	c-Myc;Esrrb;Klf4;Nanog;n-Myc;Smad1;Sox2;Stat3;
chr14:47623127-47623421	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Tcfcp2l1;
chr14:23086353-23086665	8	E2f1;Esrrb;Klf4;Nanog;n-Myc;Smad1;Sox2;Stat3;
chr14:120641446-120641694	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr13:54225492-54225800	8	E2f1;Esrrb;Klf4;Nanog;n-Myc;Sox2;Stat3;Zfx;
chr13:113583370-113583632	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Tcfcp2l1;
chr12:87393473-87393752	8	Esrrb;Klf4;Nanog;n-Myc;Smad1;Sox2;Stat3;Tcfcp2l1;
chr12:103131343-103131643	8	c-Myc;E2f1;Klf4;Nanog;n-Myc;Sox2;Stat3;Zfx;
chr11:96006351-96006538	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr11:44590264-44590536	8	Esrrb;Klf4;Nanog;Oct4;Smad1;Sox2;Stat3;Tcfcp2l1;
chr11:116787007-116787271	8	E2f1;Esrrb;Klf4;Nanog;Oct4;Sox2;Tcfcp2l1;Zfx;
chr11:102083694-102083955	8	CTCF;E2f1;Esrrb;Klf4;Nanog;Oct4;Stat3;Tcfcp2l1;
chr10:79652456-79652773	8	E2f1;Esrrb;Klf4;n-Myc;Oct4;Stat3;Tcfcp2l1;Zfx;
chr10:41155814-41156070	8	Esrrb;Klf4;Nanog;n-Myc;Oct4;Smad1;Stat3;Tcfcp2l1;
chr1:89507317-89507545	8	c-Myc;E2f1;Esrrb;Klf4;n-Myc;Stat3;Tcfcp2l1;Zfx;
chr1:84718361-84718677	8	E2f1;Esrrb;Klf4;n-Myc;Oct4;Stat3;Tcfcp2l1;Zfx;
chr1:74330516-74330882	8	c-Myc;CTCF;E2f1;Klf4;n-Myc;Stat3;Tcfcp2l1;Zfx;
chr1:7115922-7116219	8	E2f1;Esrrb;Klf4;Nanog;Smad1;Sox2;Stat3;Tcfcp2l1;
chr1:162571509-162571788	8	Esrrb;Klf4;Nanog;n-Myc;Smad1;Sox2;Tcfcp2l1;Zfx;
chr1:121132843-121133099	8	c-Myc;E2f1;Esrrb;Klf4;Nanog;Oct4;Sox2;Stat3;

Таблица П.7

Амплифицированные участки генома клеточной линии MCF-7, удаленные из статистической обработки при исследовании сайтов связывания транскрипционного фактора ERα

Номер хромосомы человека	Начало участка	Конец участка
1	105861150	107149248
1	111431534	111719134
1	113226233	114559282
3	61632702	64814111
8	86150847	129435574
15	47255525	52185855
17	45362439	67851327
17	73648047	74413579
20	39059002	41922963
20	46139677	63644868
21	38505900	46959990