

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
доктора биологических наук

Новосибирск
2014

Работа выполнена в лаборатории эволюционной биоинформатики и теоретической генетики ФГБУН Институт цитологии и генетики Сибирского отделения РАН, г. Новосибирск

Научный консультант: Колчанов Николай Александрович
академик РАН, профессор

Официальные оппоненты: Макеев Всеволод Юрьевич

Доктор физико-математических наук, зав. лабораторией системной биологии и вычислительной генетики. Федеральное государственное бюджетное учреждение науки Институт общей генетики РАН им. Н.И. Вавилова, г. Москва

Омельянчук Леонид Владимирович

Доктор биологических наук, зав. лабораторией генетики клеточного цикла. Федеральное государственное бюджетное учреждение науки Институт молекулярной биологии СО РАН, г. Новосибирск

Самсонова Мария Георгиевна

Доктор биологических наук, зав. отделом компьютерной биологии центра перспективных исследований. Санкт-Петербургский государственный политехнический университет, г. Санкт-Петербург

Ведущая организация: ФГБУН Российской академии наук Центр "Биоинженерия"
РАН, г. Москва.

Защита диссертации состоится «__» _____ 201 г. на утреннем заседании диссертационного совета по защите диссертаций на соискание ученой степени доктора наук, на соискание ученой степени кандидата наук (Д 003.011.01) при Институте цитологии и генетики СО РАН в конференц-зале по адресу: 630090, г. Новосибирск, проспект Лаврентьева, 10. Факс (383)333-12-78; e-mail: dissov@bionet.nsc.ru

С диссертацией можно ознакомиться в библиотеке Института цитологии и генетики СО РАН, и на сайте Института: www.bionet.nsc.ru

Автореферат разослан «__» _____ 201 г.

Ученый секретарь
диссертационного совета,
доктор биологических наук

Т.М. Хлебодарова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Начало XXI века ознаменовано значительными достижениями в молекулярной биологии и генетике, связанными с качественно новыми, полногеномными исследованиями. Создание высокопроизводительных методов секвенирования ДНК, измерения экспрессии генов привело к лавинообразному росту объема информации как о полных последовательностях геномов эукариот, так и о последовательностях регуляторных районов генов и о динамике экспрессии генов. Исследование регуляции экспрессии генов эукариот в масштабе генома требует изучения сайтов связывания транскрипционных факторов (ССТФ), контролирующих транскрипцию генов, установления их геномной локализации и определения генов-мишеней.

До 2005 года основным экспериментальными методами исследования ССТФ были футпринтинг ДНК, методы задержки (ретардации) в геле, ориентированные на выявление отдельных сайтов в конкретных регуляторных районах, которые не могли применяться для выявления всех сайтов связывания транскрипционного фактора в масштабе полного генома из-за чрезмерной трудоемкости и стоимости таких исследований.

Благодаря технологиям высокопроизводительного секвенирования и методам, основанным на иммунопреципитации хроматина (ChIP - Chromatin ImmunoPrecipitation), таким как ChIP-seq, ChIP-PET, ChIP-chip, появились огромные массивы качественно новых данных, позволяющих оценивать регуляторный потенциал геномов - выявлять сайты связывания транскрипционных факторов в полных геномах. Технологии олигонуклеотидных микрочипов высокой плотности и методы высокопроизводительного секвенирования транскриптомов (RNA-seq) открыли возможность идентификации групп генов, меняющих экспрессию в ответ на воздействие транскрипционных факторов и установление регуляторных взаимосвязей.

Исследование структуры хроматина на уровне отдельных нуклеосом (метилирование и ацетилирование определенных аминокислотных остатков гистонов в составе нуклеосомы) с помощью технологий ChIP-seq качественно дополняет описание регуляторных районов генов в масштабе генома. Появились экспериментальные данные о роли трехмерной организации хромосом в регуляции экспрессии генов (удаленные энхансеры, пространственные домены), полученные с помощью технологий как 3C (Chromosome Conformation Capture - определение структуры хромосомы), ChIA-PET (Chromatin Interaction

Analysis by Paired-End-Tag sequencing – анализ контактов хроматина с помощью секвенирования парных концов) и Hi-C (Баттулин и др., 2012; Li et al., 2012). Использование технологии ChIA-PET, также как и ChIP-seq, основанной на иммунопреципитации хроматина и секвенировании, позволяет экспериментально определять не только сайты связывания, но и физические контакты между удаленными районами хромосом, опосредованные белковыми комплексами. Перечисленные выше подходы ChIP-seq, ChIP-PET, ChIA-PET, использующие высокопроизводительное секвенирование, порождают огромные объемы экспериментальных данных и требуют разработки новых методов их компьютерного анализа для определения сайтов связывания транскрипционных факторов и участков модификаций хроматина, регулирующих экспрессию генов эукариот. Встает задача компьютерного описания всех ССТФ в геноме, анализа их распределения и аннотации с использованием гетерогенных информационных ресурсов по организации геномов эукариот, включая интегрированные среды хранения данных ChIP-seq, репозитории экспериментальных данных и визуализации геномной информации (Gene Expression Atlas, Ensembl, UCSC Genome Browser, GEO NCBI, ENCODE).

Отметим две актуальные задачи, в которых методы полногеномного компьютерного анализа находят большое применение. Прежде всего, это изучение транскрипционных факторов, регулирующих гены, вовлеченные в процессы канцерогенеза. Компьютерное исследование полногеномного распределения сайтов связывания таких транскрипционных факторов имеет большое значения для поиска их генов-мишеней как маркеров для диагностики. Изучение генов-мишеней онкогенов MYC (Zeller et al., 2006) и транскрипционного фактора - рецептора эстрогенов ER α (Fullwood et al., 2009; Joseph et al., 2010) в геноме человека важно для понимания молекулярных механизмов развития онкологических заболеваний, поиска новых способов их диагностики и лечения.

Вторая актуальная задача - изучение механизмов транскрипционной регуляции генов, обеспечивающих поддержание плюрипотентного состояния стволовых клеток, - имеет большое значение для исследования механизмов репрограммирования стволовых клеток (Chen et al., 2008; Yuan et al., 2009; Heng et al., 2010; Chia et al., 2010; Орлов и др., 2012). Эта задача включает определение полногеномных карт сайтов связывания регуляторов плюрипотентности - транскрипционных факторов NANOG, OCT4, SOX2, KLF4, PRDM14 в эмбриональных стволовых клетках (ЭСК). Один из ключевых подходов к решению этой

задачи - полногеномный компьютерный анализ распределения ССТФ в геномах на основе данных ChIP-seq.

Представляемая диссертационная работа посвящена применению современных компьютерных методов анализа регуляции транскрипции эукариот с использованием данных экспериментов высокопроизводительного секвенирования и иммунопреципитации хроматина.

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 c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2 и Smad2 в геноме мыши. Реконструкция распределения сайтов связывания транскрипционных факторов MYC, PRDM14, ER α , FOXA1, OCT4, NANOG в геноме человека.
3. Компьютерное исследование ассоциации сайтов связывания транскрипционного фактора ER α 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 ER α 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 ER α 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 компьютерные методы анализа распределения сайтов связывания транскрипционных факторов в геноме на основе данных ChIP-seq (Kuznetsov et al., 2007; Orlov et al., 2012). С помощью этих программ построены карты связывания транскрипционных факторов с-Мус, Oct4, Nanog, Sox2, E2f1, n-Мус, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши (Chen et al., 2008; Yuan et al., 2009; Han et al., 2010), фактора Zic3 в геноме рыбы *Danio rerio* (Winata et al., 2013), транскрипционных факторов MYC (Zeller et al., 2006), ER α , FOXA1 (Joseph et al., 2010), PRDM14 в геноме человека (Chia et al., 2010). Представленные 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 (Chen et al., 2008; Orlov et al., 2012; Орлов и др., 2012). Разработанная компьютерная база данных наборов проб микрочипов платформы Affymetrix U133 была новой на момент публикации (Orlov et al., 2007), и использовалась для анализа присутствия транскриптов в цис-антисенс ориентации (Grinchuk et al., 2010).

Компьютерный анализ по данным ChIP-seq впервые показал статистически значимую совместную локализацию сайтов связывания транскрипционных факторов Oct4, Sox2, Nanog, с одной стороны и с-Мус, n-Мус с другой, в эмбриональных стволовых клетках (ЭСК) мыши (Chen et al., 2008; Han et al., 2010; Heng et al., 2010).

Впервые рассчитана совместная локализация сайтов связывания транскрипционных факторов плюрипотентности в геномах мыши и человека в форме матриц сближенности

ССТФ (Heng et al., 2010). Определен новый нуклеотидный мотив сайта связывания транскрипционного фактора PRDM14 в геноме человека (Chia et al., 2010).

Впервые построен компьютерный метод предсказания сайтов связывания ТФ ER α 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 (Joseph et al., 2010). С помощью разработанных автором компьютерных программ впервые совместно с данными ChIP-seq проанализированы карты хромосомных контактов, опосредованных связыванием белка рецептора эстрогенов ER α (Fullwood et al., 2009), полученные посредством технологии ChIA-PET.

Впервые на основе компьютерного анализа полногеномных данных о хромосомных контактах, опосредованных комплексами РНК-полимеразы II, данных транскрипционной активности генов, и профилей модификаций гистонов для пяти клеточных линий в геноме человека (Li et al., 2012) показана положительная корреляция участков хромосомных контактов с модификациями гистонов, характеризующими открытое состояние хроматина (H3K4me3, H3K9ac, H3K4me1).

Теоретическая значимость работы. Разработана компьютерная статистическая модель распределения сайтов связывания транскрипционных факторов, позволяющая 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 ER α в геноме человека за счет одновременного анализа, как нуклеотидных последовательностей, так и профилей модификаций 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 компьютерных методов состоит в возможности поиска регуляторных районов генов по данным секвенирования в масштабе полного генома эукариот. Программный комплекс ICGenomics (<http://www->

bionet.sccc.ru/icgenomics/) для функциональной аннотации геномных последовательностей обеспечивает существенное расширение методов компьютерного анализа полногеномных данных. Созданы база данных качества наборов проб микрочипов Affymetrix U133 (Orlov et al., 2007a; 2007b), база цис-антисенс транскриптов в геноме человека (Grinchuk et al., 2010) и база данных экспрессии генов на микрочипах для крыс RatDNA (<http://pixie.bionet.nsc.ru/ratdna/rat/index.php>; свидетельство госрегистрации №621051 от 10.10.2012г).

По тематике исследования 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 (№№ 11-04-01771, 11-04-01888, 11-04-92712-ИНД_а, 12-04-00897), Интеграционные проекты СО РАН (№ 119, 136), проект № 8740 Минобрнауки России «Интегрированная биоинформационная платформа анализа данных экспрессии генов в тканях мозга». Продолжается работа по грантам РФФИ № 14-04-01906 и 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-Мус, n-Мус 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 ER α в геноме человека положительно ассоциировано с районами метилирования и ацетилирования гистонов нуклеосом H3K4me3, H3K4me1, H3K9ac и H3K14ac. Разработан компьютерный алгоритм для предсказания сайтов связывания ER α 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 «компьютерный и теоретический анализ данных, статистическая обработка». Для получения результатов, представленных в диссертации, автором были разработаны алгоритмы и компьютерные программы анализа данных ChIP-seq, статистического моделирования, сравнения геномных координат и геномной аннотации, оценки качества наборов проб микрочипов, анализа нуклеотидных контекстов, компьютерной симуляции полногеномных данных ChIP-PET, ChIP-seq и ChIA-PET.

Методика анализа полноты эксперимента ChIP-seq представлена в работе (Orlov et al., 2009), 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). В статье (Joseph, Orlov et al., 2010) посвященной исследованию сайтов связывания транскрипционного фактора ER α в геноме человека автор диссертации является автором для переписки.

Основные результаты применения разработанных компьютерных методов для анализа распределений сайтов связывания транскрипционных факторов представлены в

статьях, написанных в соавторстве. В статьях (Chen, ... Orlov et al., 2008) и (Han, ... Orlov et al., 2010) автор диссертации выполнил компьютерный анализ данных и оценил полноту эксперимента ChIP-seq в ЭСК мыши. Программа анализа профиля ChIP-seq, разработанная автором, использовалась в исследованиях транскрипционных факторов, опубликованных в статьях (Yuan, ... Orlov et al., 2009; Heng, ... Orlov et al., 2010; Lee, ... Orlov et al., 2011). В статье (Chia, ... Orlov et al., 2010) автором диссертации выполнен анализ CCTФ PRDM14 и компьютерная реконструкция генных сетей плюрипотентности в ЭСК человека. В статье (Zhao, ... Orlov et al., 2007) с помощью разработанных компьютерных моделей исследованы полногеномные данные модификаций гистонов H3K4me3 и H3K27me3. Обобщение анализа распределений CCTФ в ЭСК человека и мыши дано в статье (Heng, Orlov, Ng, 2010). В статье (Winata, ... Orlov et al., 2013) представлено применение разработанных автором программ для анализа CCTФ в геноме *D. rerio*. Публикация (Zeller, ... Orlov et al., 2006) показывает применение компьютерных моделей анализа эксперимента ChIP-PET для TF с-Мус. В работе (Goh, Orlov et al. 2010) с помощью разработанной автором компьютерной программы исследован профиль плотности нуклеосом в геноме дрожжей. В статьях (Fullwood, ... Orlov et al., 2009) и (Li, ..., Orlov et al., 2012) 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; Пугта, Орлов и соавт., 2011), исследованию экспрессии генов на микрочипах: (Orlov et al. 2007a; Orlov et al., 2007b; Орлов и соавт., 2011; Кожевникова, ... Орлов, 2012; Kozhevnikova, ... Orlov et al. 2013; Медведева, ... Орлов, 2013).

Апробация работы. Результаты были представлены на Пятой, Шестой, Седьмой и 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), 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 (С.-Петербург), съезде-конференции ВОГиС-2013 (Новосибирск), MCCMB-2013 (Москва), конференции «Нейроинформатика-2014» (Москва).

Публикации. По теме диссертации опубликовано 33 печатные работы, из них 30 – статьи в научных изданиях (журналы по списку ВАК).

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», и Приложения. 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 этих факторов, анализ распределения сайтов связывания транскрипционного фактора – рецептора эстрогенов ER α . 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 человека и мыши, показана роль транскрипционных факторов в поддержания плюрипотентности и репрограммировании. Дан обзор проблем исследования трехмерных контактов хромосом в ядре с помощью секвенирования (методы 3C, Hi-C и ChIA-PET).

Глава 2 содержит описание компьютерных моделей распределения ССТФ в эукариотическом геноме. Представлены алгоритмы анализа данных ChIP-seq и базы

микрочиповых данных, разработанные автором (Kuznetsov et al., 2007; Orlov et al., 2009).

Глава 3 посвящена применению разработанных автором программ к реконструкции карт сайтов связывания транскрипционных факторов, полученных по экспериментальным данным ChIP-seq (Zeller et al., 2006; Chen et al., 2008; Yuan et al., 2009; Han et al., 2010; Chia et al., 2010; Lee et al., 2011; Winata et al., 2013).

Глава 4 содержит описание применения разработанных компьютерных методов к исследованию нуклеосомной упаковки и модификаций хроматина в геноме человека, полученных с помощью ChIP-seq, анализ моделей сайтов связывания ER α (Fullwood et al., 2009; Joseph et al., 2010; Orlov et al., 2012; Орлов и др., 2012). Показано применение программ для анализа связывания транскрипционных факторов в геноме дрожжей (Goh et al., 2010).

Глава 5 представляет исследование хромосомных контактов, полученных с помощью массового параллельного секвенирования в геноме человека по методу ChIA-PET для ER α и комплекса РНК-полимеразы II с помощью разработанных автором диссертации компьютерных программ (Fullwood et al., 2009; Li et al., 2012; Баттулин и др., 2012).

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

Благодарности. Автор выражает глубокую признательность научному консультанту академику РАН Н.А. Колчанову, сотрудникам ИЦиГ СО РАН Д.А. Афонникову и В.А. Иванисенко за поддержку и обсуждение научных результатов, В.А. Кузнецову и соавторам из Геномного Института Сингапура за научную дискуссию и техническую поддержку на ранних этапах работы.

Глава 2. МОДЕЛИ РАСПРЕДЕЛЕНИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ В ГЕНОМЕ

Базы данных и выборки геномных последовательностей. Использовались следующие данные: (1) первичные данные секвенирования участков ДНК, полученные методом ChIP-seq для факторов c-Мус, Oct4, Nanog, Sox2, E2f1, n-Мус, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши, транскрипционных факторов MYC, FOXA1, ER α , PRDM14 в геноме человека; (2) данные по экспрессии генов на микрочипах платформы Affymetrix U133 в опухолевых клетках 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 методами ChIP-PET (GSE18046 ER α человека), ChIP-seq и ChIA-PET. Исходные данные секвенирования ChIP-seq представлены архивами GSE11431 для факторов Nanog, Oct4, STAT3, Smad1, Sox2, Zfx, c-Мус, n-Мус, Klf4, Esrrb, Tcfcp2l1, E2f1, CTCF, p300 и Suz12, GSE19219 для Tbx3, GSE17439 и GSE17642 для Eset, GSE19019 для Nr5a2, GSE23581 для Smad2 ЭСК мыши, GSE26831 и GSE23893 для ER α человека (культуры клеток MCF-7), GSE22767 и GSE22792 для PRDM14 (ЭСК человека), GSE41458 для Zic3 рыбы *D.rerio*. Секвенирование нуклеосомной ДНК дрожжей *S.cerevisiae* представлено архивом GSE26392. Данные ChIA-PET представлены архивами GSE18046 для ТФ ER α , и GSE33664 – для РНК-полимеразы II человека.

Алгоритмы анализа профилей связывания ChIP-seq в масштабе генома. Метод иммунопреципитации хроматина ChIP-seq основан на обработке клеток формальдегидом, вызывающим образование ковалентных сшивок между ДНК и белками. Хроматин дробится на отдельные фрагменты с помощью ультразвука; далее с помощью иммунопреципитации со специфическими антителами выделяются последовательности ДНК, физически связанные с исследуемым белком. На следующем этапе выполняется очистка ДНК с последующим секвенированием всего набора выделенных фрагментов на приборах массового параллельного секвенирования ДНК (по технологиям Roche 454, Illumina Solexa или SOLiD).

Первичные данные секвенирования ДНК (прочтения, или «риды» ChIP-seq, размером от 18 до 75 нуклеотидов) конвертируются в линейную карту позиций прочтений по хромосомам в геноме, образуя профиль ChIP-seq связывания исследуемого белкового

фактора с ДНК (рис. 1). Алгоритм определения пиков полногеномного профиля, составленного из фрагментов ДНК, и статистического выделения значимых пиков, был реализован первоначально для метода ChIP-PET (здесь PET обозначает парные концы - Paired End Tags) (Zeller et al., 2006), и далее для метода ChIP-seq (Chen et al., 2008).

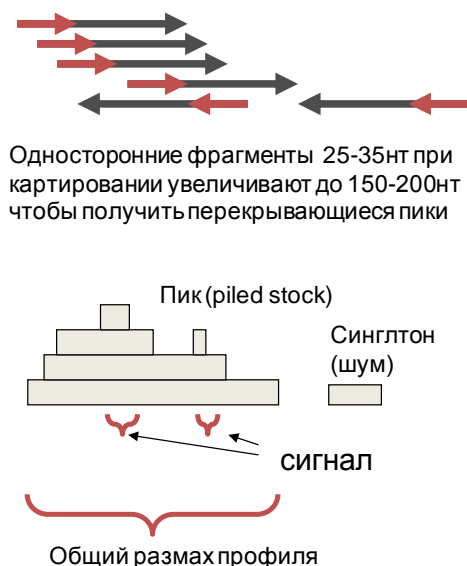


Рис. 1. Определение кластера (пика) сформированного односторонними ChIP фрагментами прочтений ДНК в хромосомных координатах.

Дальнейшая модификация алгоритма поиска пиков была выполнена для метода ChIA-PET (Fullwood et al., 2009), учитывающего уже не только отдельные сайты, но и пары пространственно сближенных сайтов на хромосомах.

Координаты ChIP прочтений удлиняются на размер фрагмента (150-200 нт), строится ступенчатый профиль фрагментов наложенных друг на друга в геномных координатах. Затем определяются пики такого геномного профиля. Пик – наиболее высокая точка профиля, более вероятно содержащая сайт связывания транскрипционного фактора (рис. 1).

Задача определения сайтов связывания (набора неслучайных пиков в геноме) решается статистически с помощью сравнения экспериментально полученного распределения набора пиков по высоте с распределением, ожидаемым по случайным причинам (рис. 2). Автором предложен компьютерный алгоритм определения пиков профиля и их последующей фильтрации от «шума» и ошибок секвенирования, основанный на статистике распределения числа сайтов в геноме в зависимости от высоты пика ChIP-seq (Kuznetsov et al., 2007; Orlov et al., 2009).

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

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

где P_{ob} - функция вероятности наблюдаемого распределения встречаемости ChIP пиков, X – пики, $m=1,2,3,...$ – высота пика, P_{sp} – вероятность специфических пиков, $0 < \alpha < 1$ – доля специфических пиков в общем распределении пиков в геноме, P_{ns} – вероятность

неспецифичных пиков (шумовой сигнал) в общем распределении пиков по профилю в геноме (Orlov et al., 2009).

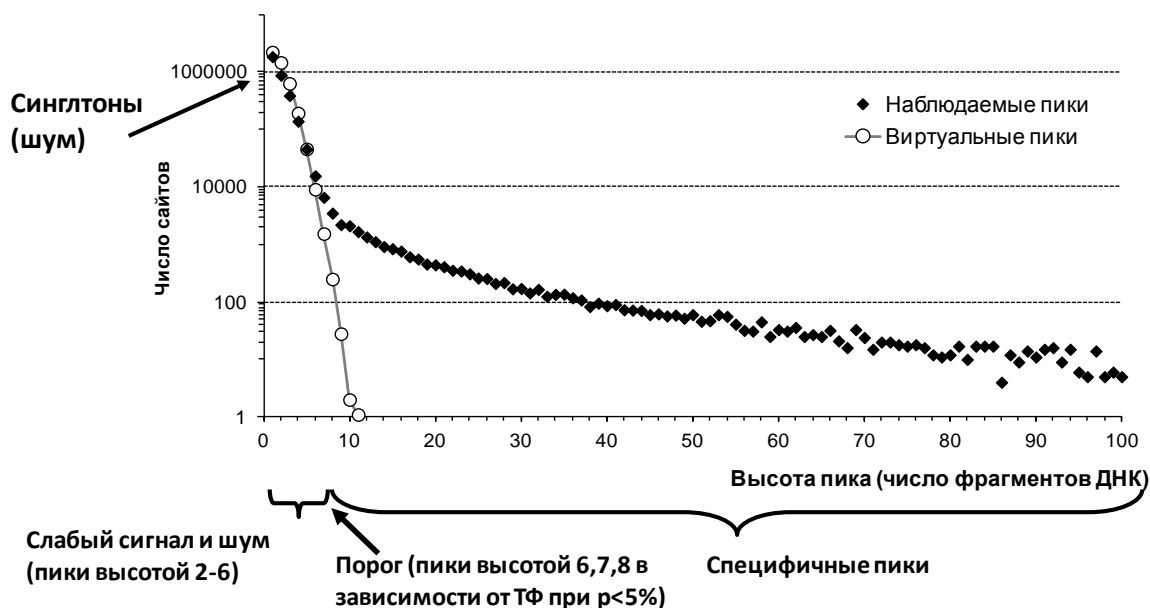


Рис. 2. Распределение числа сайтов в геноме в зависимости от высоты пика ChIP-seq (для связывания ТФ Nanog мыши) (Chen et al., 2008; Orlov et al., 2009).

На основе экспериментальных ChIP данных строится эмпирическая функция распределения пиков. Распределение числа специфичных пиков P_{sp} в зависимости от их высоты может быть оценено с помощью распределения Парето, начиная с большой высоты пика, где шумового сигнала уже нет (например, с $m=10$ на рис. 2).

Неспецифичное распределение P_{ns} может быть оценено с помощью компьютерной симуляции по появлению кластеров (пиков) при случайном (равномерном) распределении прочтений ДНК вдоль хромосомы (Kuznetsov et al., 2007).

Метод оценки полноты (насыщения) эксперимента ChIP-seq по определению ССТФ в масштабе генома основан на компьютерных симуляциях и экстраполировании числа определяемых пиков как функции числа прочтений ДНК (глубины секвенирования) в эксперименте (рис. 3) (Chen et al., 2008, Orlov et al., 2009).

На рисунке 3 представлен пример зависимости роста порогового значения высоты пика и числа специфичных пиков от числа прочтений (размера библиотеки ChIP-seq). Для компьютерной симуляции в качестве исходных данных использовались геномные координаты прочтений в эксперименте ChIP-seq. Заданная доля прочтений удалялась с помощью компьютерной программы, использующей генератор случайных чисел. Затем

выполнялся новый расчет пиков в геноме при уменьшенном размере библиотеки, и процедура повторялась итеративно (Chen et al., 2008).

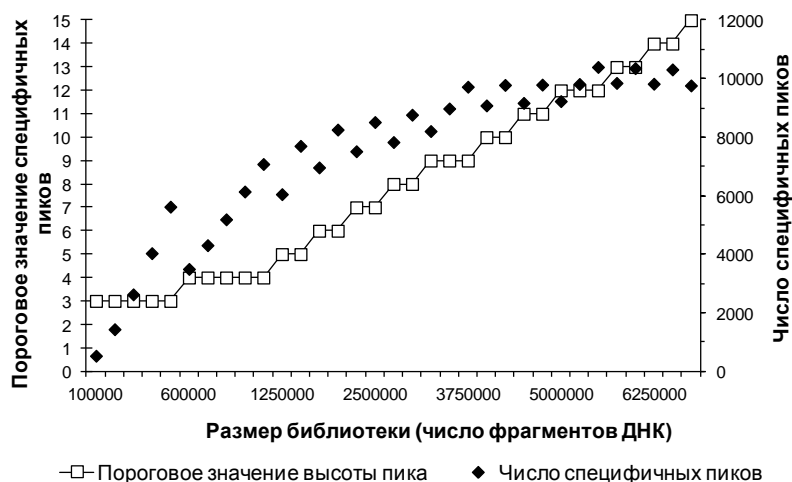


Рис. 3. Зависимость порогового значения высоты пика (ось Y слева) и числа выявленных пиков (ось Y справа) в геноме в зависимости от глубины секвенирования в рассматриваемом эксперименте ChIP-seq (ТФ Nanog мыши) (Orlov et al., 2009).

объема данных секвенирования. При увеличении глубины секвенирования повышаются пороги распознавания, как на истинные пики, так и на ложные сайты.

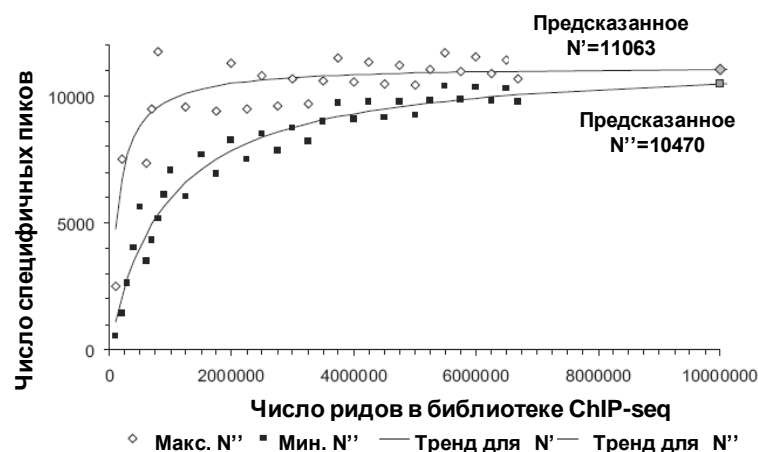


Рис. 4. Анализ сатурации (полноты) определения ССТФ в геноме в заданном эксперименте ChIP-seq (Orlov et al., 2009).

формуле:

$$f(X) = B_{max} * X / (K + X) \quad (2)$$

где B_{max} — максимальное число сайтов (при насыщении эксперимента), K — константа

На каждом шаге рассчитывалось пороговое значение и число найденных пиков в геноме.

На рисунке 4 представлен расчет предсказания общего числа сайтов в геноме (оценки верхней и нижней границ такого числа) для эксперимента ChIP-seq по связыванию ТФ Nanog с помощью экстраполяции

Когда при экстраполяции вперед число определяемых в геноме сайтов перестает расти, эксперимент ChIP-seq можно считать полным. Определение полноты выполненного эксперимента $f(X)$ как пропорции числа фактически найденных сайтов в геноме, определялось по

нормализации, X – число прочтений (объем данных секвенирования).

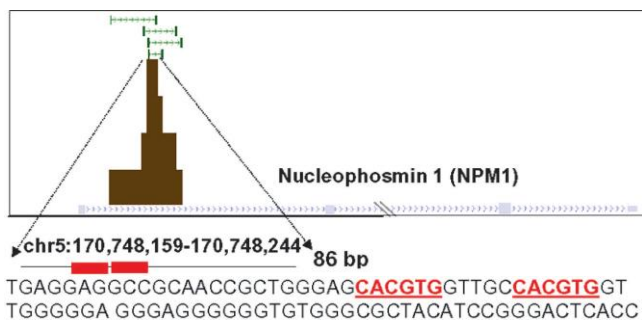


Рис. 5. Определение мотива связывания Мус в нуклеотидных последовательностях профиля ChIP-PET. Показан известный сайт связывания в первом интроне гена NPM1 (Zeller et al., 2006).

Предложенный метод был применен для анализа 15 наборов экспериментов ChIP-seq в ЭСК мыши. Было показано, что 5-6 миллионов фрагментов уже достаточно, чтобы выявить все сайты с достаточной (90%-95%) точностью (Chen et al., 2008).

Поиск мотивов сайтов связывания в нуклеотидных последовательностях, соответствующих пикам профилей

ChIP-seq, далее выполнялся с помощью собственных программ и с использованием программ MEME, Weeder, STAMP, баз данных TRANSFAC и JASPAR (рис. 5). Алгоритм определения генов-мишеней транскрипционных факторов по данным экспрессии генов на микрочипах, разработанный автором, был основан на определении расстояния от ССТФ до старта транскрипции ближайшего гена, аннотированного в геноме.

Важным этапом предобработки данных экспрессии на микрочипах является проверка качества и фильтрация данных измерения сигнала. Распространенная платформа микрочипов Affymetrix U133 имеет недостатки дизайна целевых (таргетных) последовательностей наборов проб, приводящие к систематическим ошибкам, как показано в (Orlov et al., 2007a; Orlov et al., 2007b). Для каждой целевой последовательности набора проб микрочипа Affymetrix U133 автором было рассчитано пересечение с геномными повторами, аннотированными в RepBase, по семействам и типам (DNA, LTR, LINE, SINE включая MIR и Alu). Разработанные программы применялись для анализа сигнала проб на микрочипе на клинических выборках данных экспрессии генов в клетках опухолей молочной железы и глиом мозга. Показана статистически значимая ассоциация шумового сигнала экспрессии наборов проб микрочипа Affymetrix со сходством нуклеотидных последовательностей этих наборов проб к мобильным элементам генома человека (Orlov et al., 2007a; Орлов и др., 2011).

Глава 3. КАРТЫ САЙТОВ СВЯЗЫВАНИЯ ПО ДАННЫМ ChIP-seq

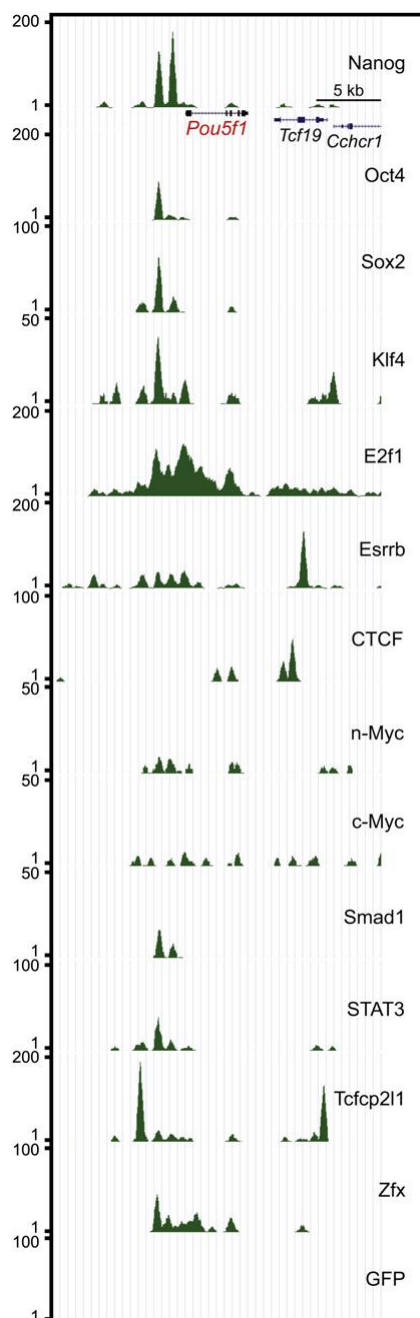


Рис. 6. Профили связывания 13 различных транскрипционных факторов в геноме мыши для гена *Pou5f1* (Chen et al., 2008). По оси абсцисс – положение на хромосоме. По оси ординат - высота профиля ChIP-seq (число прочтений ДНК).

Распределение сайтов связывания транскрипционных факторов по данным ChIP-seq исследовалось с помощью разработанных компьютерных программ определения пиков полногеномных профилей. На ЭСК мыши были проведены эксперименты ChIP-seq для ТФ Nanog, Oct4, Sox2, Klf4, E2f1, Esrrb, CTCF, n-Myc, c-Myc, Smad1, STAT3, Tcfcp2l1, Zfx, Suz12, а также контрольное секвенирование (с белком GFP) (Chen et al., 2008).

Показана возможность реконструкции регуляторных контуров ТФ на основе расположения сайтов связывания в промоторах генов-мишеней и данных экспрессии генов на микрочипах. Установлено, что ключевые факторы плюрипотентности Oct4, Nanog, Sox2 образуют тесно связанную регуляторную сеть (Chen et al., 2008).

Серия исследований ChIP-seq на том же типе клеток была продолжена экспериментами по определению сайтов связывания факторов Eset, Nr5a2, Tbx3, Smad2, представленными в работах автора (Yuan et al., 2009; Heng et al., 2009; Han et al., 2010; Lee et al., 2011).

Используя данные ChIP-seq для профилей связывания ТФ в ЭСК мыши, далее исследовались локусы множественного связывания в геномных районах, размером до 500 нт (так называемые, множественные локусы регуляции транскрипции). Учитывая размер генома, число неперекрывающихся промоторных районов и размер сайта, рассчитано, что кластеры, состоящие из 4 и более сайтов, могут рассматриваться как неслучайные ($p < 0.01$) (рис.7).

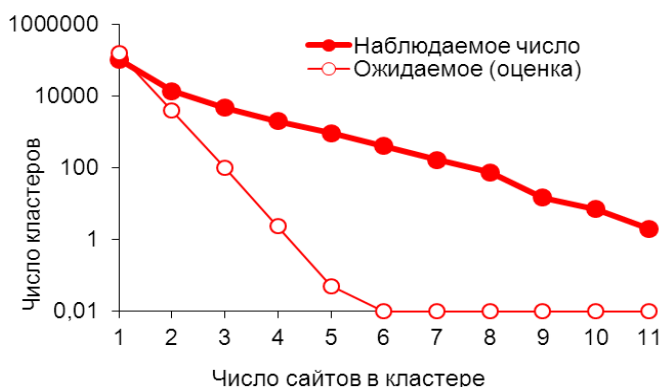


Рис. 7. Оценка встречаемости кластеров сайтов связывания ТФ в геноме мыши (Chen et al., 2008).

сайтов связывания с помощью разработанной компьютерной программы позволяет численно оценить совместную локализацию регуляторных районов различных транскрипционных факторов в геноме.

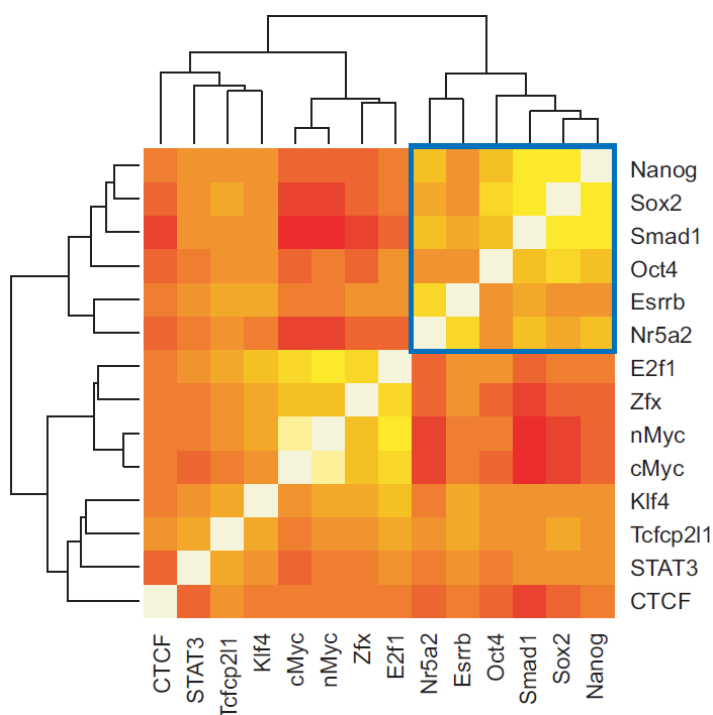


Рис. 8. Термокарта ко-локализации ССТФ в геноме мыши, включая сайты связывания факторов Oct4 - Nanog - Sox2 - Klf4, а также Nr5a2 (Yuan et al., 2009).

факторов, ответственных за поддержание плюрипотентности: выделен кластер, включающий ТФ из базового набора факторов репрограммирования Oct4-Nanog-Sox2-Klf4 (Chen et al., 2008). Показана ко-локализация сайтов связывания факторов n-Мус, с-Мус, Zfx и E2f1.

Расчет матрицы сближенности расположения сайтов в геноме, используемой для построения тепловой карты, проводился с помощью собственной компьютерной программы, последовательно рассчитывающей для каждой двух независимых экспериментов ChIP-seq число сайтов, расположенных друг от друга не более чем на 100 нт. Расчет таких кластеров

Кластеризация сайтов связывания различных факторов по взаимному расположению (рис. 8) в геноме мыши представлена на термокарте (тепловой карте), исходя из объединенных данных нескольких экспериментов (Chen et al., 2008; Yuan et al., 2009; Heng et al., 2010; Han et al., 2010). Более яркий цвет в ячейке термокарты соответствует более частой совместной локализации сайтов пары факторов, указанных справа и внизу для данной ячейки. Показана совместная локализация (ко-локализация) сайтов связывания

ключевых факторов плюрипотентности, включающих набор OCT4-NANOG-SOX2, что и для ЭСК мыши (Chia et al., 2010; Heng et al., 2010).

Компьютерный анализ показал совместную локализацию сайтов связывания PRDM14 с регуляторными районами транскрипционных факторов плюрипотентности, формируемыми сайтами связывания OCT4, SOX2, NANOG (Chia et al., 2010)..

Дозовый эффект связывания ТФ. Компьютерное исследование эффекта связывания ТФ в зависимости от дозы белка активатора Activin, ингибитора SB431542, а также измерение экспрессии генов на микрочипах выполнялось для CCTФ Smad2 в ЭСК мыши (Lee et al., 2011). Для трех состояний обработки клеток в ЭСК мыши (после обработки активатором белком Activin, растворителем DMSO и ингибитором SB431542) были доступны данные ChIP-seq по связыванию ДНК с белком Smad2, а также данные измерений уровней экспрессии генов на микрочипах. Для каждого состояния активации был построен свой профиль ChIP-seq, определены пики и определены гены-мишени по расположению относительно гена (рис. 11).

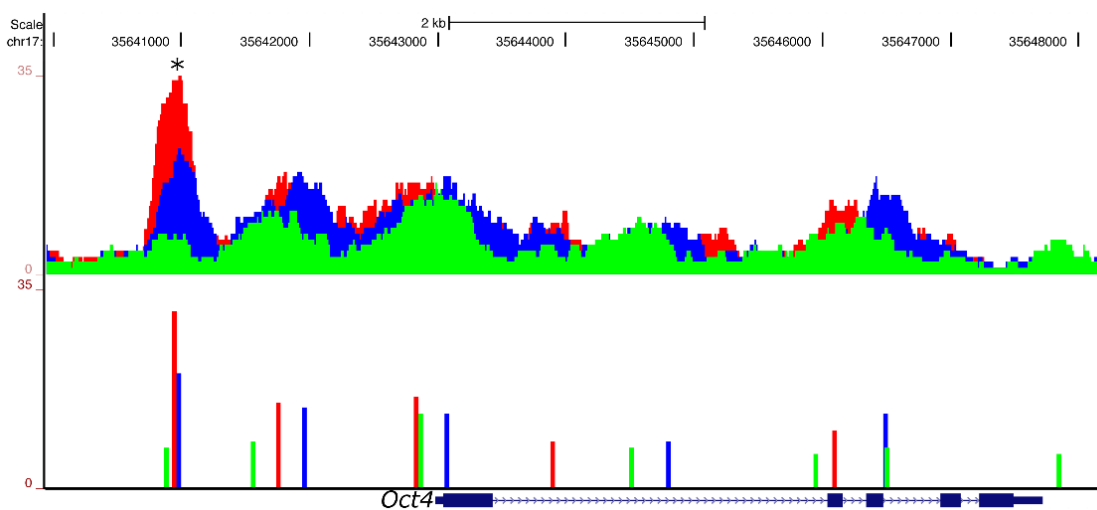


Рис. 11. Профили связывания Smad2 в трех условиях и поиск мотивов CCTФ в пике профиля ChIP-seq (Lee et al., 2011). Цветовые обозначения профилей ChIP-seq и пиков – активация белком активин Activin (красный), нейтральная среда - DMSO (синий) и ингибитор SB431542 (зеленый). По оси абсцисс – положение профилей на хромосоме 17 мыши, по оси ординат – высота профиля ChIP-seq (число прочтений ДНК). На нижней панели показаны позиции выделенных сайтов Smad2.

В этом проведенном впервые компьютерном тройном сравнении ChIP-seq для одного ТФ выявлены разные группы сайтов связывания Smad2 и их гены-мишени (Lee et al., 2011). Показано, что увеличение/уменьшение уровня сигнала нодал/активин (Nodal/Activin) ведет к

транскрипционной регуляции специфических групп мишеней Smad2 в ЭСК мыши. С помощью компьютерного анализа нуклеотидных последовательностей в пиках профилей ChIP-seq, кроме мотива связывания Smad2, установлено присутствие вторичных мотивов других ТФ (ко-мотивов), влияющих на выбор генов-мишеней Smad2 (Lee et al., 2011).

В целом, глава 3 содержит результаты применения разработанных компьютерных методов анализа профилей ChIP-seq к исследованию ССТФ в эмбриональных стволовых клетках в геноме человека и в геноме мыши (Chen et al., 2008; Chia et al., 2010; Heng et al., 2010; Lee et al., 2011). Применение разработанных автором методов анализа данных ChIP-seq показано для транскрипционного фактора Zic3 для рыбы *D.rerio* (Winata et al., 2013). Для большинства исследованных транскрипционных факторов показана высокая доля сайтов связывания, расположенных в дистальных регуляторных районах.

Глава 4. МОДИФИКАЦИИ ХРОМАТИНА И СВЯЗЫВАНИЕ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ В ГЕНОМЕ ПО ДАННЫМ ChIP-seq

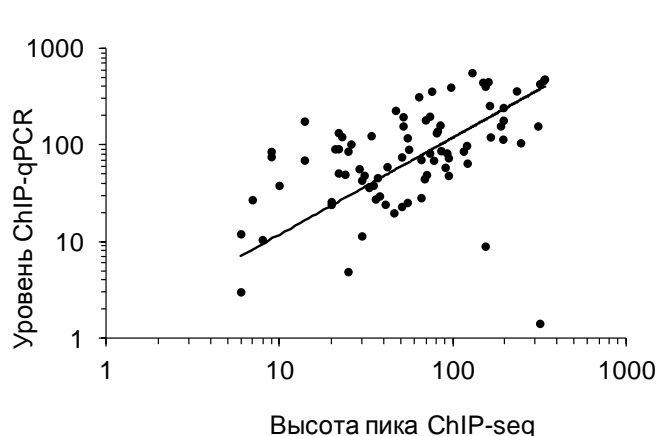


Рис. 12. Экспериментальная проверка 81 сайта связывания ERα с помощью ChIP-qPCR (количественная ПЦР ChIP-фрагментов ДНК) (Joseph et al., 2010).

Связывание белковых факторов транскрипции с ДНК определяется состоянием хроматина, которое также можно исследовать в масштабе генома с помощью технологий иммунопреципитации хроматина и последующего компьютерного анализа. Встает задача создания предсказательной компьютерной модели связывания с белком на основе маркеров состояния хроматина. Эта задача

решена на примере компьютерного анализа данных ChIP-seq для транскрипционного фактора - рецептора эстрогенов ERα и модификаций гистонов, также определенных с помощью ChIP-seq, в клеточных линиях MCF-7 и T47D (Joseph et al., 2010).

Для фактора ERα был определен нуклеотидный мотив и уточнена весовая матрица связывания (Joseph et al., 2010). Показана корреляция между высотой пика геномного

профиля ChIP-seq и силой связывания, измеренной для 81 сайта ER α с помощью ChIP-qPCR, в двойной логарифмической шкале (рис. 12), что подтверждает корректность теоретической модели распределения пиков. Линейный коэффициент корреляции составил 0.56 (уровень значимости $P=5.0E-8$), ранговый коэффициент корреляции Кендалла 0.375 ($P=7.17E-07$).

Целью исследования было построение модели определения (предсказания) реально связанных сайтов в геноме в данном типе клеток, с помощью дополнительной геномной информации, используя аннотированные экспериментальные данные по состоянию хроматина (модификациям гистонов). Использовались 4 типа метилирования гистона H3 (по позициям лизина) – H3K4me1, H3K4me3, H3K9me3, H3K27me3, и 2 типа модификаций ацетилирования – H3K9ac, H3K14ac (Joseph et al., 2010). Состояние доступности ДНК для белков может быть измерено с помощью метода FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements), оценивающего степень доступности ДНК для связывания с белками. Использовались данные нескольких независимых экспериментов ChIP-seq в двух условиях: в нейтральной среде и после обработки клеток эстрадиолом (E2), активирующим экспрессию ER α . Показана статистическая обогащенность сайтами связывания ER α промоторов генов, увеличивающих экспрессию после обработки E2 (Joseph et al., 2010).

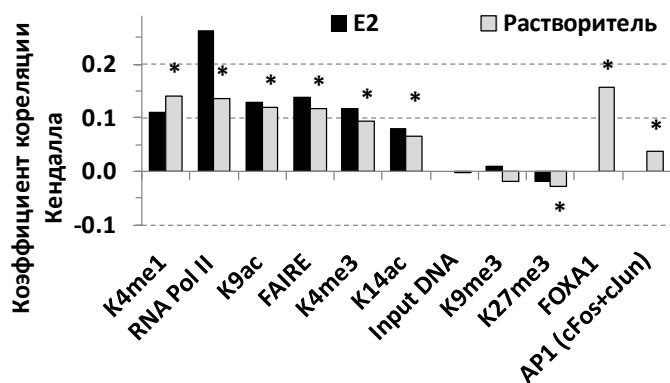


Рис. 13. Ранговый коэффициент корреляции профилей ChIP-seq присутствия модификаций хроматина со связыванием ER α в геноме человека (Joseph et al., 2010).

Показана позитивная ассоциация участков открытого хроматина в геноме с сайтами связывания ER α (темные красные столбики для ChIP-seq маркеров модификаций гистонов) (рис. 13). Статистическое подтверждение ассоциации между связыванием ТФ и изменением структуры хроматина (доступности) было получено для маркеров активации

транскрипции (модификации лизина в соответствующих позициях гистона H3: ацетилирование H3K9ac, H3K14ac, метилирование H3K4me1, H3K4me3). В то же время нет корреляции с модификациями H3K9me3 и H3K27me3, соответствующими закрытому хроматину. Интенсивность связывания коррелирует и с профилем секвенирования модификаций хроматина в нейтральной среде до активации, и с профилем связывания

фактора FOXA1, который меняет состояние хроматина (Joseph et al., 2010).

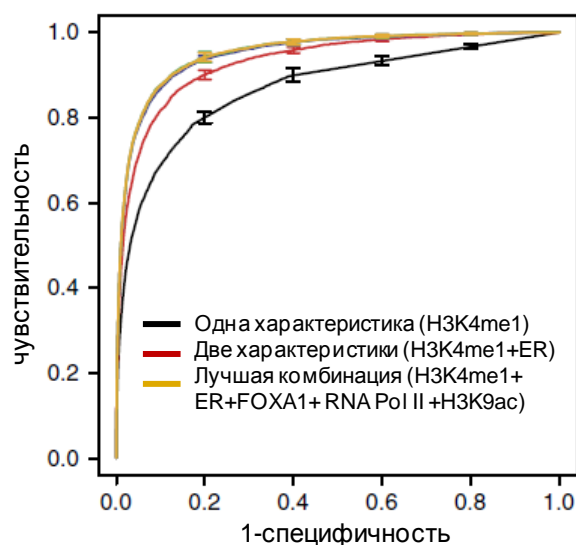


Рис. 14. Кривая ошибок (ROC) для предсказания сайтов связывания ER α в геноме человека по данным секвенирования модификаций гистонов ChIP-seq (Joseph et al., 2010).

Оценки зависимостей от состояния хроматина были использованы для создания модели предсказания сайтов связывания ER α в геноме, только по профилям модификаций гистонов (рис. 14). По обучающей выборке сайтов связывания и «не-сайтов», т.е. участков, не имеющих сродства к ER α , и данным о состоянии хроматина по ChIP-seq была построена модель логистической регрессии для предсказания связывания ER α . Точность этой модели оказалась достаточно высокой (до 0.95 по площади под кривой ошибок ROC-AUC) (рис. 14).

С помощью анализа данных ChIP-seq по связыванию ER α в геноме для культур клеток MCF-7 и T47D была показана тканеспецифичность сайтов связывания этого транскрипционного фактора (Joseph et al., 2010).

В целом, глава 4 содержит описание применения разработанных компьютерных методов к исследованию модификаций хроматина в геноме, анализ моделей поиска и предсказания связывания для сайтов ER α , FOXA1 (Fullwood et al., 2009; Goh et al., 2010; Joseph et al., 2010; Orlov et al., 2012; Орлов и др., 2012).

Глава 5. ХРОМОСОМНЫЕ КОНТАКТЫ И РЕГУЛЯЦИЯ ТРАНСКРИПЦИИ В ГЕНОМЕ ЧЕЛОВЕКА

Проблема анализа трехмерной структуры генома активно исследуется в последние годы с помощью методов, основанных на секвенировании. 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 (Fullwood et al., 2009; Li et al., 2012).

В этом методе, так же как для ChIP-seq, сначала выполняется иммунопреципитация хроматиновых комплексов, связанных с исследуемым белком, затем фрагменты ДНК разделяются ультразвуком (рис. 15).

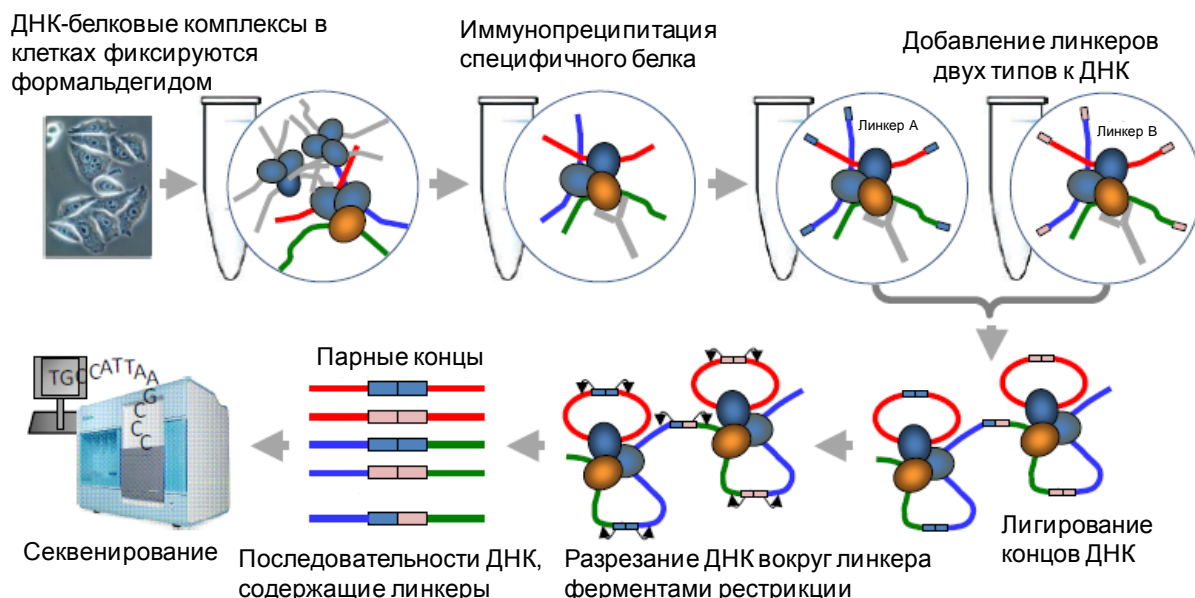


Рис. 15. Схема метода ChIA-PET (Fullwood et al., 2009). Синим, красным и зеленым цветом обозначены последовательности ДНК, контактирующие с белками, которые затем лигируются. Парные концы вместе с линкером направляются на секвенирование.

Фрагменты ДНК из разделенных ультразвуком, прошедших иммунопреципитацию хроматиновых комплексов обрабатываются через лигирование линкеров (на свободные концы ДНК), получают 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.

На следующем рисунке показана ассоциация контактирующих сайтов связывания ER α с комплексом РНК-полимеразы II, модификацией гистонов H3K4me3 и сайтами FOXA1. Сравнение полногеномных экспериментов, основанных на иммунопреципитации хроматина ChIP-PET, ChIP-seq и ChIA-PET для сайтов связывания ER α в культуре клеток

MCF-7 показало устойчивость определения сайтов с помощью этих технологий (Fullwood et al., 2009) (рис.16).

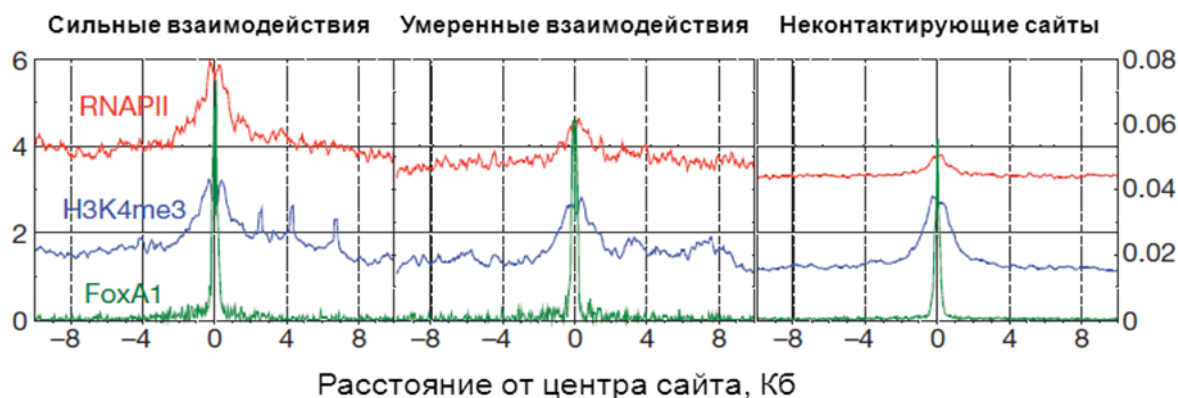


Рис. 16. Ассоциация сайтов связывания ERα, определенных с помощью ChIA-PET (Fullwood et al., 2009) различной интенсивности (три панели слева направо – от сильных взаимодействий до отсутствия контактов) с профилями РНК-полимеразы II, модификацией H3K4me3 (по данным ChIP-seq, левая ось Y) и сайтами FOXA1 (данные ChIP-on-chip, правая ось Y) (Fullwood et al., 2009).

Присутствие маркера открытого хроматина H3K4me3 для сайтов ERα, определенных с помощью ChIP-seq, может использоваться для предсказания сайтов в геноме, также как и присутствие сигнала связывания FOXA1 (Joseph et al., 2010).

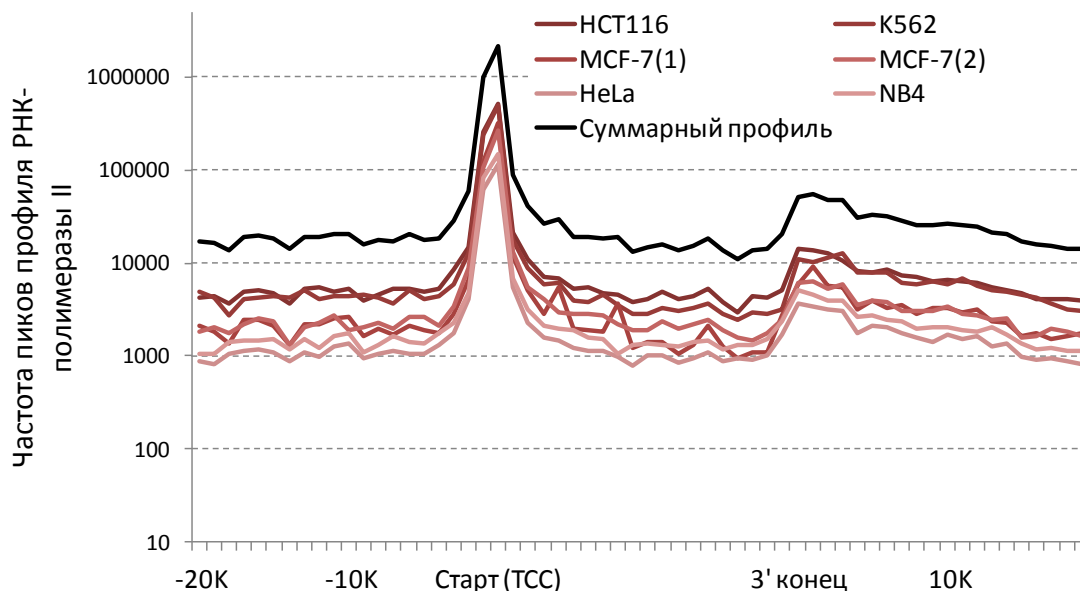


Рис. 17. Профили связывания РНК-полимеразы II в экспериментах ChIA-PET относительно генов в культурах клеток человека HCT116, K562, MCF-7, HeLa и NB4 (Li et al., 2012).

Профиль связывания РНК-полимеразы II (рис. 17) относительно генов (на основе аннотации баз данных RefSeq и UCSC genes) для экспериментов в культурах клеток человека построен по участкам связывания в геноме, определенных методом ChIA-PET (Li et al., 2012).

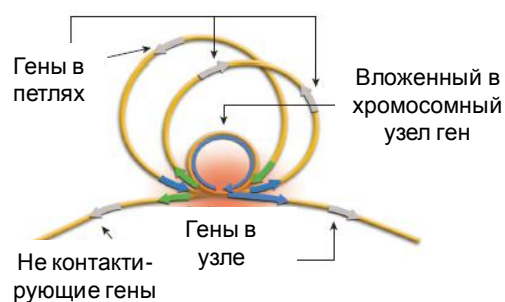


Рис. 18. Модельная структура петель хромосом, содержащих контактирующие гены (Fullwood et al., 2009).

Такое распределение с пиками в районе старта транскрипции и 3' района указывает на возможность образования петли между 5'- и 3'-районами гена. Более того, несколько генов могут быть включены в петли хромосом, контактируя друг с другом промоторными районами, либо 3'-районами (рис. 18). Показана обогатненность контактирующих участков, опосредованных комплексом РНК-полимеразы

II, сайтами связывания различных транскрипционных факторов, полученными в экспериментах ChIP-seq.

Классификация моделей промоторных, энхансерных и мультигенных контактов, опосредованных комплексом РНК-полимеразы II (Li et al., 2012) приведена на рисунке 19.

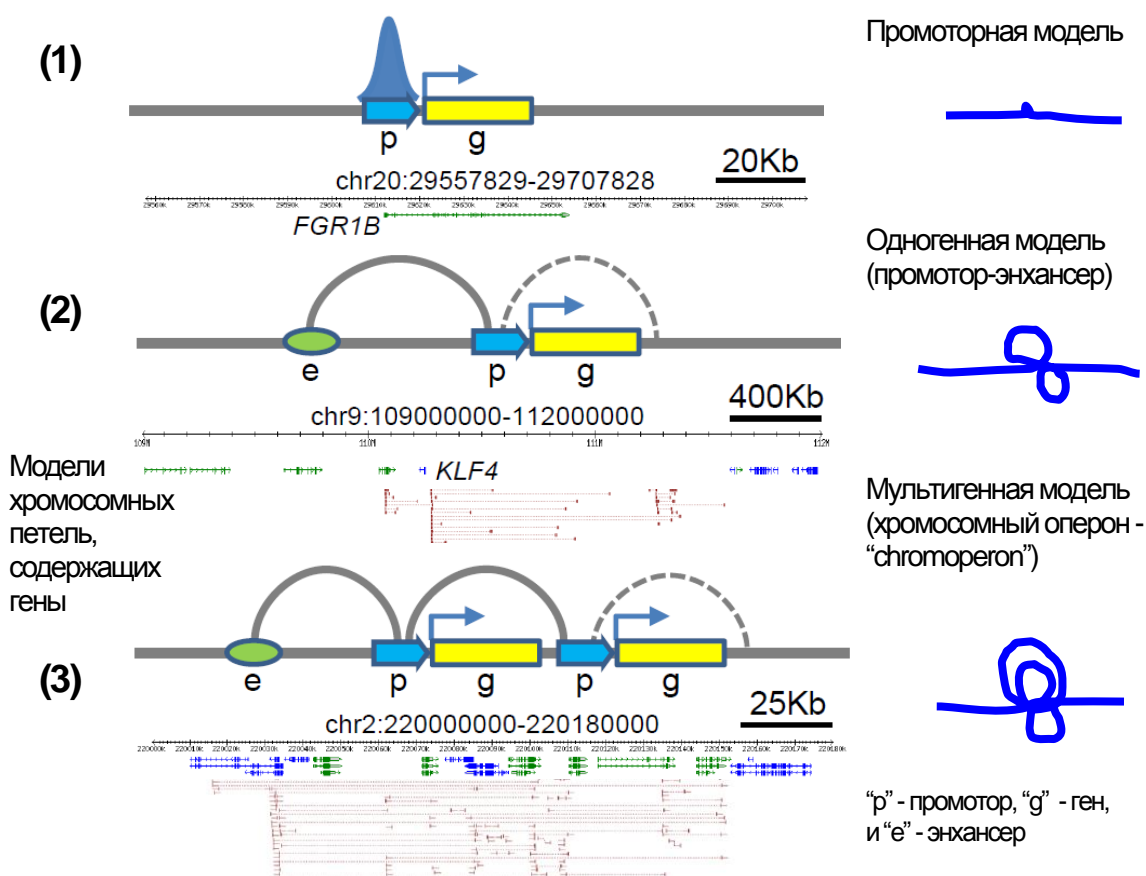


Рис. 19. ChIA-PET. Модели промоторных (1), энхансерных (2) и мультигенных (3) хромосомных контактов, опосредованных комплексом РНК-полимеразы II (Li et al., 2012).

Модель базального промотора включает только локальные петли ДНК в промоторе,

без удаленных взаимодействий. Одногенная модель включает только петли в районе гена между энхансером и промотором, возможно между 5' и 3' районами гена, но без других белок-кодирующих генов. И наконец, мультигенная модель включает сразу несколько генов, расположенных рядом друг с другом и контактирующих промоторными районами. В этой модели также возможен контакт промоторов с удаленными энхансерами.

Как было показано в предыдущих главах работы, важнейшее значение для регуляции транскрипции генов имеют модификации хроматина, прежде всего гистона H3, модификации лизина в позициях 4, 9, 14, включающие метилирование и ацетилирования, связанные с доступностью ДНК для связывания белковых факторов транскрипции (Zhao et al., 2007). Исследование контактов хромосом, опосредованных комплексом РНК-полимеразы II в культурах клеток человека, подтвердило ассоциацию контактирующих районов с такими маркерами модификаций хроматина (Li et al., 2012). На рисунке 20 показаны профили модификаций хроматина для участка хромосомы 2 человека, содержащий ген *PCBP1*.

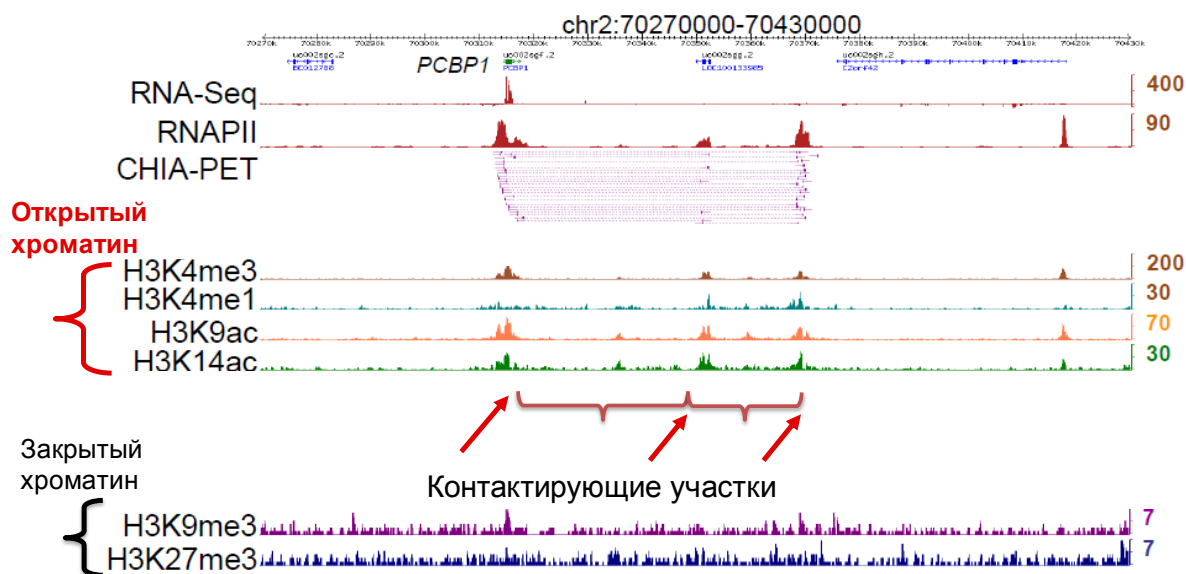


Рис. 20. Пример профилей ChIP-seq модификаций гистонов, соответствующих открытому и закрытому состоянию хроматина (группы показаны слева), в участках хромосомных контактов, найденных с помощью ChIA-PET (отмечено внизу) (Li et al., 2012).

Рисунок 20 показывает, что хроматин в контактирующих участках открыт: гистоны имеют маркеры модификации активной транскрипции H3K4me3, H3K9ac (видны пики профиля); в то же время модификации репрессии транскрипции H3K9me3, H3K27me3 не имеют пиков (равномерный шум).

Более того, исследованные ассоциации хромосомных контактов, опосредованных

белком ER α , также связаны с активацией хроматина, в частности с модификацией H3K4me3, как показано для генома человека методом ChIA-PET (Fullwood et al., 2009).

В целом, Глава 5 представляет компьютерное исследование хромосомных контактов, полученных с помощью массового параллельного секвенирования в геноме человека по методу ChIA-PET (Fullwood et al., 2009; Li et al., 2012; Баттулин и др., 2012). Показана ассоциация участков хромосомных контактов, опосредованных комплексом РНК-полимеразы II с сайтами связывания ТФ, определенными в экспериментах ChIP-seq (Li et al., 2012).

Заключение

В рамках диссертационной работы разработаны компьютерные программы анализа распределения сайтов связывания транскрипционных факторов в геноме на основе данных экспериментов по иммунопреципитации хроматина ChIP-seq, ChIP-chip, ChIP-PET. Впервые представлена компьютерная модель оценки насыщения (полноты) эксперимента ChIP-seq с помощью симулирования распределения пиков в геноме и экстраполяции на большее число прочтений ДНК. Разработана компьютерная база данных качества проб на микрочипах Affymetrix (Orlov et al., 2007a; Orlov et al., 2007b).

Обработан большой набор полногеномных карт связывания транскрипционных факторов, включая c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши (Chen et al., 2008; Yuan et al., 2009; Han et al., 2010) и ССТФ MYC (Zeller et al., 2006), ER α (Joseph et al., 2010), PRDM14 в геноме человека (Chia et al., 2010). Показана кластеризация сайтов связывания транскрипционных факторов Oct4, Sox2, Nanog в геноме ЭСК мыши, подтвержденная последующими данными ChIP-seq, что свидетельствует о связности регуляторной сети поддержания плюрипотентности (Chen et al., 2008; Han et al., 2010; Heng et al., 2010).

Компьютерное исследование ассоциации сайтов связывания ТФ ER α с модификациями гистонов показало возможность точного полногеномного предсказания сайтов на основе данных ChIP-seq о модификациях хроматина (Joseph et al., 2010). Анализ полногеномных данных ChIA-PET о хромосомных контактах, опосредованных ER α и комплексом РНК-полимеразы II, позволил предложить мультигенную модель регуляции транскрипции, и статистически описать удаленные регуляторные районы генов человека на хромосомах (Fullwood et al., 2009; Li et al., 2012).

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*.

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 ER α 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 ER α , в культурах клеток опухолей человека P493 и MCF-7, соответственно. Выявлены нуклеотидные мотивы транскрипционных факторов, связывающихся в окрестностях сайтов ER α .

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 ER α в геноме человека.

10) Впервые выполнен компьютерный анализ карт хромосомных контактов, опосредованных рецептором эстрогенов ER α 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 основных публикаций по теме диссертации

Статьи в научных журналах

1. Орлов Ю.Л., Левицкий В.Г., Смирнова О.Г., Подколотная О.А., Хлебодарова Т.М., Колчанов Н.А. Статистический анализ последовательностей ДНК, содержащих сайты формирования нуклеосом // *Биофизика*. – 2006 – Т. 51. – С. 608-14.

2. 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. – Т. 51. – С. 602-7.
3. **Orlov Y.L.**, Te Boekhorst R., Abnizova I.I. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information // *J Bioinform Comput Biol*. – 2006. – V.4. – P.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. 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. – P.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. 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. – N. 3. – P.286-98.
6. **Orlov Y.L.**, Zhou J., Lipovich L., Shahab A., Kuznetsov V.A. Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis // *In Silico Biol*. – 2007a. – V.7. – N. 3. – P.241-60.
7. Kuznetsov V.A., **Orlov Y.L.**, Wei C.L., Ruan Y. Computational analysis and modeling of genome-scale avidity distribution of transcription factor binding sites in chip-pet experiments // *Genome Inform*. – 2007. – V.19. – P.83-94.
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. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells // *Cell*. – 2008. – V. 133. – N. 6. – P. 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. An oestrogen-receptor-alpha-bound human chromatin interactome // *Nature*. – 2009. – V. 462. – N. 7269. – P. 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. Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells // *Genes Dev.* – 2009. – V.23. – N. 21. – P. 2507-20.
11. Grinchuk O.V., Jenjaroenpun P., **Orlov Y.L.**, Zhou J., Kuznetsov V.A. Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns // *Nucleic Acids Res.* – 2010. – V. 38. – N. 2. – P. 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. The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells // *Cell Stem Cell.* – 2010. – V. 6. – N. 2. – P. 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. Tbx3 improves the germ-line competency of induced pluripotent stem cells // *Nature.* – 2010. – V. 463. – N. 7284. – P. 1096-100.
14. Goh W.S., **Orlov Y.**, Li J., Clarke N.D. 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. – N. 1. – P. e1000649.
15. Guo X., Popadin K.Y., Markuzon N., **Orlov Y.L.**, Kraysberg 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. – V. 26. – N. 8. – P. 340-3.
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. A genome-wide RNAi screen identifies PRDM14 as a regulator of POU5F1 and human embryonic stem cell identity // *Nature.* – 2010. – V. 468. – N. 7321. – P. 316-20.
17. Heng J.C., **Orlov Y.L.**, Ng H.H. Transcription Factors for the Modulation of Pluripotency and Reprogramming // In: *Cold Spring Harb Symp Quant Biol.* – 2010. – V. 75. – P. 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. Integrative model of genomic factors for determining binding site selection by estrogen receptor α . // *Mol Syst Biol.* – 2010. – V. 6. – P.456.
19. **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. – Т. 15. – № 2. – С. 327-339.

20. Lee K.L., Lim S.K., **Orlov Y.L.**, Yit le Y., Yang H., Ang L.T., Poellinger L., Lim B. Graded Nodal/Activin signaling titrates conversion of quantitative phospho-Smad2 levels into qualitative embryonic stem cell fate decisions // *PLoS Genet.* – 2011. – V. 7. – N. 6. – P. e1002130.
21. Пугта П., **Орлов Ю.Л.**, Подколотный Н.Л., Митра Ч.К. Относительно консервативные общие короткие последовательности в сайтах связывания транскрипционных факторов и мРНК // *Вавиловский журнал генетики и селекции.* – 2011. – Т.15. – № 4. – С. 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. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation // *Cell.* – 2012. – V.148. – N.1-2. – P.84-98.
23. 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. – Т. 16. – №4/1. – С. 756-765.
24. **Орлов Ю.Л.**, Брагин А.О., Медведева И.В., Гунбин И.В., Деменков П.С., Вишневский О.В., Левицкий В.Г., Ощепков В.Г., Подколотный В.Г., Афонников В.Г., Гроссе И., Колчанов Н.А. ICGenomics: программный комплекс анализа символьных последовательностей геномики // *Вавиловский журнал генетики и селекции.* – 2012. – Т. 16. – №4/1. – С. 732-741.
25. **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. 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. – N. 2. – P. 211.
26. Баттулин Н.Р., Фишман 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. – Т. 16. – № 4/2. – С. 872-878.
27. Winata C.L., Kondrychyn I., Kumar V., Srinivasan K.G., **Orlov Y.L.**, Ravishankar A., Prabhakar S., Stanton L.W., Korzh V., Mathavan S. Genome-wide analysis reveals Zic3 interaction with distal regulatory elements of stage specific developmental genes in zebrafish // *PLOS Genetics* – 2013. – V. 9(10). – e1003852.

28. Kozhevnikova O.S., Korbolina E.E., Stefanova N.A., Muraleva N.A., **Orlov Y.L.**, Kolosova N.G. Association of AMD-like retinopathy development with an Alzheimer's disease metabolic pathway in OXYS rats // *Biogerontology*. – 2013. – 14(6). – 753-62.
29. Медведева И.В., Вишнеvский 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. – Т. 17. – № 4/1. – С. 629-638.
30. **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.

Статьи в сборниках

31. **Orlov Y.L.**, Zhou 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* (J.C. Ragapakse, B. Schmidt and G. Volkert, Eds), Springer-Verlag: Berlin-Heidelberg – 2007b. P. 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. Genome-wide statistical analysis of multiple transcription factor binding sites obtained by ChIP-seq technologies // In: *Proceedings of the 16th ACM Workshop on Breaking Frontiers of Computational Biology (CompBio '09)*. ACM, New York, NY. – 2009. – P. 11-18.
33. Kolchanov N.A., **Orlov Y.L.** Introductory note for BGRS-2012 special issue // *Journal of Bioinformatics and Computational Biology*. – 2013. – V. 11. – N. 1. – P. 1302001.

Подписано к печати « » 201 г.
Формат бумаги 60 x 90 1/16. Печ. л. 2. Уч.изд.л. 1,4
Тираж 110 экз. Заказ № .

Ротапринт Института цитологии и генетики СО РАН
630090, Новосибирск, проспект академика Лаврентьева, 10.