

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ

ЦЕНТР «БИОИНЖЕНЕРИЯ» РОССИЙСКОЙ АКАДЕМИИ НАУК

Юридический адрес: 117312 Москва:
просп. 60-летия Октября, 7, корп. 1
ИНН 7728118561/ КПП 772801001

Тел. (499) 135-73-19
Факс (499) 135-05-71
office@biengi.ac.ru

УТВЕРЖДАЮ

Заместитель директора по научной работе
Центра "Биоинженерия" РАН, д.б.н.
Н.В. Равин
«10» ноября 2014 г.

ОТЗЫВ

ведущей организации - Центра "Биоинженерия" РАН - на диссертационную работу Орлова Юрия Львовича «Полногеномный компьютерный анализ распределения сайтов связывания транскрипционных факторов эукариот по данным иммунопреципитации хроматина и высокопроизводительного секвенирования», представленную на соискание ученой степени доктора биологических наук по специальности 03.01.09 - Математическая биология, биоинформатика.

Диссертационная работа Ю.Л. Орлова посвящена разработке компьютерных методов исследования регуляторных районов генов эукариот с помощью полногеномного анализа сайтов связывания транскрипционных факторов, что относится к фундаментальной проблеме молекулярной биологии - исследованию механизмов контроля экспрессии генов эукариот.

Актуальность темы. В диссертации поставлена задача компьютерного определения участков связывания транскрипционных факторов эукариот в масштабе генома, на основе анализа данных, получаемых с помощью технологий иммунопреципитации хроматина и высокопроизводительного секвенирования ДНК (ChIP-seq в английской аббревиатуре). С практической точки зрения понимание регуляторных процессов в масштабе генома позволяет решать различные задачи биоинженерии, биотехнологии и биомедицины.

В работе рассмотрены задачи предсказания положения регуляторных районов на основе данных ChIP-seq и анализ положения дистальных регуляторных районов генов,

определяемых хромосомными контактами. Вопросы компьютерного анализа регуляции экспрессии генов в масштабе генома эукариот требуют внимания исследователей и остаются технически сложными, поскольку продолжающееся быстрое развитие высокопроизводительных технологий секвенирования ДНК, методов экспериментального определения участков связывания транскрипционных факторов опережает развитие компьютерных методов анализа таких полногеномных данных. Предлагаемый набор компьютерных решений для анализа данных ChIP-seq, а также смежных технологий, использующих иммунопреципитацию хроматина - ChIP-PET и ChIA-PET, соответствует возникающим требованиям науки.

Отметим две большие 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 участков связывания таких транскрипционных факторов имеет большое значения для поиска их генов-мишеней как маркеров для диагностики заболеваний. Изучение генов-мишеней онкогенов MYC и транскрипционного фактора рецептора эстрогенов ER α в геноме человека важно для понимания молекулярных механизмов развития онкологических заболеваний, поиска новых способов их диагностики и лечения.

Другая актуальная задача - изучение механизмов транскрипционной регуляции генов, обеспечивающих поддержание плюрипотентного состояния стволовых клеток, - имеет большое значение для исследования механизмов соматического репрограммирования стволовых клеток. Эта задача включает определение полногеномных карт участков связывания регуляторов плюрипотентности - транскрипционных факторов NANOG, OCT4, SOX2, KLF4 и других в эмбриональных стволовых клетках (ЭСК). Подход к решению этой важнейшей задачи состоит в компьютерном анализ распределения сайтов связывания в масштабе генома на основе данных ChIP-seq.

Рассмотренные компьютерные решения находят применение для анализа распределения участков связывания транскрипционных факторов в масштабе генома. Рассмотрены применения для анализа участков связывания транскрипционного фактора MYC, транскрипционного фактора - рецептора эстрогенов в геноме человека. Рассчитаны позиции положения в геноме участков связывания ряда транскрипционных факторов, ответственных за поддержание плюрипотентности в эмбриональных стволовых клетках (Nanog, Oct4, Sox2 и ряда других) в геноме мыши и в геноме человека. Для генома мыши по данным ChIP-seq рассчитаны позиции участков связывания и определены кластеры таких участков для более 15

различных транскрипционных факторов.

Диссертационная работа представляет ряд смежных проблем, возникающих при компьютерном моделировании положения участков транскрипционного фактора в эукариотическом геноме, в частности, связанных с определением генов-мишеней действия этого фактора, ко-локализации с другими регуляторными районами генов, реконструкции генной сети регуляторных взаимодействий. Важна компьютерная интеграция разработанных программных средств обработки данных ChIP-seq с программами анализа последовательностей ДНК, определения нуклеотидных мотивов, предсказания положения сайтов связывания, успешно реализованная в данной работе, и отраженная в научных публикациях автора, представленных в диссертации. Вышесказанное еще раз подчеркивает актуальность диссертационной работы.

Научная новизна исследования, полученных результатов, выводов и рекомендаций, сформулированных в диссертации.

В работе представлены оригинальные компьютерные методы анализа распределения участков связывания транскрипционных факторов в геноме на основе данных иммунопреципитации хроматина, сопряженной с последующим высокопроизводительным секвенированием фрагментов ДНК (ChIP-seq).

Построены карты связывания транскрипционных факторов c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши, транскрипционных факторов MYC, ER α , FOXA1, PRDM14 в геноме человека. Карта участков связывания для инсуляторного фактора CCTF получена впервые.

Разработана компьютерная база данных наборов проб микрочипов платформы Affymetrix U133, которая была новой на момент публикации, и использовалась далее для анализа присутствия транскриптов в цис-антисенс ориентации в геноме человека и создании соответствующей базы данных перекрывающихся транскриптов.

Компьютерный анализ по данным ChIP-seq впервые показал совместную локализацию участков связывания транскрипционных факторов Oct4, Sox2, Nanog, с одной стороны и c-Myc, n-Myc с другой, в эмбриональных стволовых клетках (ЭСК) мыши. Совместная локализация участков связывания транскрипционных факторов плюрипотентности в геномах мыши и человека впервые была представлена в форме матриц сближенности и тепловых карт в публикациях автора. Выполнен анализ нуклеотидных мотивов de novo, для участков связывания факторов Zic3 в геноме *D.rerio*, Smad2 в геноме мыши. Определен новый нуклеотидный состав участка связывания транскрипционного фактора PRDM14 в геноме человека.

Впервые построен компьютерный метод предсказания участков связывания транскрипционного фактора - рецептора эстрогенов ER α в масштабе генома на основе профилей модификации хроматина - ацетилирования и метилирования гистонов (модификации H3K4me3, H3K4me1, H3K27me3, H3K9me3, H3K9ac, H3K14ac), определенных с помощью технологии ChIP-seq в клеточных линиях MCF-7 и T47D. С помощью разработанных компьютерных программ впервые совместно с данными ChIP-seq проанализированы карты хромосомных контактов, полученные посредством технологии ChIA-PET, для рецептора эстрогенов ER α . Сам компьютерный анализ данных современной технологии ChIA-PET, основанной на иммунопреципитации контактирующих хромосомных участков с помощью секвенирования парных концов (PET), также сделан впервые.

На основе компьютерного анализа полногеномных данных о хромосомных контактах, опосредованных комплексами РНК-полимеразы II, данных транскрипционной активности генов, и профилей модификаций гистонов в геноме человека показана положительная корреляция расположения геномных участков хромосомных контактов с модификациями гистонов, характеризующими открытое состояние хроматина (H3K4me3, H3K9ac, H3K4me1 и ряда других), что расширяет теоретические представления о транскрипционной активности генов и механизмах регуляции их экспрессии.

Значимость полученных соискателем результатов для науки и практической деятельности.

Теоретическая значимость работы, связана с разработкой компьютерной статистической модели распределения участков связывания транскрипционных факторов, позволяющей определять их локализацию по координатам секвенированных «чтений» в геноме и оценивать полноту эксперимента ChIP-seq. Построена компьютерная модель, обеспечивающая высокую точность предсказания связывания 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, образованных регуляторными районами транскрипции генов (промоторов и энхансеров) в геноме человека. Такая модель основана данных технологии ChIA-PET по хромосомным контактам, опосредованным комплексом РНК-полимеразы II.

Научно-практическая ценность разработанных компьютерных методов состоит в возможности поиска регуляторных районов генов по данным секвенирования в масштабе полного генома эукариот. Представлен программный комплекс ICGenomics для

функциональной аннотации геномных последовательностей, который обеспечивает существенное расширение методов компьютерного анализа полногеномных данных. В работе представлены база данных качества наборов проб микрочипов Affymetrix U133, база цис-антисенс транскриптов в геноме человека, база данных экспрессии генов на микрочипах для крыс RatDNA, для которой представлено свидетельство госрегистрации баз данных и программ ЭВМ.

По тематике исследования выполнен ряд государственных контрактов (заказчик - Министерство образования и науки РФ) на разработку программного обеспечения для геномных исследований, грантов РФФИ, что свидетельствует о практической направленности и значимости работы.

Структура и содержание работы.

Структура диссертации образована пятью главами: «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. Первая глава представляет обзор современного состояния исследований в мире и ставит задачи исследования; вторая глава описывает разработку методов компьютерного анализа данных ChIP-seq, и родственных технологий, анализа экспрессии генов. Третья, четвертая и пятая главы описывают применение разработанных средств анализа участков связывания транскрипционных факторов в эмбриональных стволовых клетках человека и мыши, построение полногеномных карт этих факторов, анализ геномного распределения рецептора эстрогенов. В четвертой главе анализ CCTF 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.

Первая Глава, «Обзор литературы», содержит информацию о современных исследованиях регуляторных районах транскрипции в геноме человека, включая определение участков связывания с помощью технологий иммунопреципитации хроматина (ChIP). Представлены исследования по регуляции экспрессии генов, связанных с образованием опухолей, таких как рецептор эстрогенов, 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 методов исследования трехмерных контактов хромосом в

ядре с помощью секвенирования (3C, Hi-C и ChIA-PET).

Вторая Глава содержит описание компьютерных моделей распределения ССТФ в эукариотическом геноме. Представлены алгоритмы анализа данных ChIP-seq и базы микрочиповых данных, разработанные автором.

Третья Глава показывает применение разработанных программ к реконструкции карт сайтов связывания транскрипционных факторов, полученных по экспериментальным данным ChIP-PET и ChIP-seq.

Четвертая Глава содержит описание применения разработанных компьютерных методов к исследованию нуклеосомной упаковки и модификаций хроматина, полученных с помощью ChIP-seq, анализ моделей участков связывания рецептора эстрогенов. Показано применение программ для анализа связывания транскрипционных факторов в геноме дрожжей.

Пятая Глава представляет компьютерное исследование хромосомных контактов, полученных с помощью массового параллельного секвенирования в геноме человека по методу ChIA-PET, и опосредованных связыванием белка ER α и комплекса РНК-полимеразы II. Приложение содержит таблицы координат сайтов в геноме, рассчитанные по данным ChIP-seq, результаты анализа кластеризации сайтов, описание использованных компьютерных ресурсов биоинформатики.

Работа основана на большом объеме исходных данных, примеров и расчетов. Результаты представлены в высокорейтинговых научных публикациях. По каждой главе диссертации и работе в целом сделаны четкие выводы.

Рекомендации по использованию результатов и выводов диссертационной работы.

Результаты работы используются в курсе биоинформатики для студентов Новосибирского госуниверситета. Разработанные методы анализа полноты эксперимента ChIP-seq с помощью компьютерной симуляции числа участков связывания исследуемого транскрипционного фактора в зависимости от глубины секвенирования могут быть использованы при планировании и анализе экспериментов, основанных на иммунопреципитации хроматина и высокопроизводительном секвенировании ДНК.

Метод предсказания участков связывания транскрипционного фактора - рецептора эстрогенов по профилям модификаций гистонов может быть адаптирован для предсказания положения и других транскрипционных факторов. Наборы участков связывания с ДНК рецептора эстрогенов могут быть использоваться для определения генов-мишеней этого транскрипционного фактора. Компьютерные базы данных аннотации наборов проб микрочипов Affymetrix могут быть использованы для препроцессинга микрочиповых данных.

Данные о положении в геноме человека участков хромосомных контактов, опосредованных комплексом РНК-полимеразы II, могут быть использованы для анализа дистальных регуляторных районов, предсказания положения новых транскрипционных факторов в геноме.

Рекомендуется использовать полученные результаты в дальнейших исследованиях в научных учреждениях РАН, РАМН, Минобрнауки и Минздрава России, где ведется изучение взаимодействий ДНК с белками, исследования с помощью технологий высокопроизводительного секвенирования ДНК, таких как МГУ, НИЦ «Курчатовский институт», Центр «Биоинженерия» РАН, Институт общей генетики РАН им. Н.И. Вавилова, Институт молекулярной биологии им. Энгельгардта РАН, Институт физико-химической медицины МЗ РФ, Институт биохимии им. Баха РАН, Российский Онкологический научный центр им. Блохина РАМН, Институт биоорганической химии им. Шемякина и Овчинникова РАН.

Практические рекомендации, сформулированные в диссертации, обоснованы проведенными исследованиями и могут служить руководством в работе. Большое число публикаций в рекомендованных ВАК Минобрнауки РФ изданиях и в высокорейтинговых журналах (более 30 публикаций) с достаточной полнотой отражает содержание работы и позволяет оценить ее новизну, практическую и научную значимость. Включая тезисы конференций, общее число публикаций по теме диссертации 52.

Личный вклад автора диссертации состоит в компьютерном анализе данных и статистической обработке. Все представленные в диссертации результаты получены лично автором. Им 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.

Апробация результатов диссертационной работы осуществлялась в виде стендовых и устных докладов на российских и международных научных конференциях. Результаты работы были представлены на серии международных конференций по Биоинформатике и Регуляции Структуры Генома (BGRS'06, BGRS'08, BGRS\SB-2010, BGRS\SB-2012 и BGRS\SB-2014 - Новосибирск), конференциях HUGO (2008, Хайдарабад, Индия; 2010, Монпелье, Франция; 2013, Сингапур), CSHL-UK – 2007 (Великобритания), Международном Симпозиуме по Биотехнологии (Москва, 2011), Школе BREW-2011 (Тарту, Эстония), конференциях Постгеном-2011 и Постгеном-2012, Конференции Integrative Bioinformatics-2012 (Ханчжоу, Китай), SysPatho-2012 (Санкт-Петербург), конференциях ВОГиС-2013 (Новосибирск),

МССМВ-2013 (Москва), «Нейроинформатика-2014» (Москва).

Замечания к работе.

В целом, результаты, полученные автором, являются новыми научными знаниями биоинформатики.

Однако, по представленным публикациям очень сложно говорить о роли автора в получении и интерпретации экспериментальных данных. Современная геномика является синтезом работы большого коллектива авторов. Эта работа начинается от появления гипотезы и продолжается в экспериментальном лабораторном поиске, за которым следует математическая обработка данных, с последующим пониманием биологических смыслов и валидацией полногеномных данных. Очевидно, что роль Соискателя в первую очередь состояла в профессиональном анализе больших объемов экспериментальных данных. Лишь в трех англоязычных печатных работах (не принимая во внимание тезисы конференций) автор был так называемым «первым автором», т.е. его вклад в работу был ключевым среди всех авторов, и ни разу Соискатель не был «последним автором», что говорит, что он не был инициатором и идеологическим руководителем проведенных работ. Следует, однако, отметить, что современная геномика всегда есть синтез эксперимента и математической обработки, и порой не всегда удастся четко определить, кто сделал более важную работу - биоинформатики или экспериментаторы. В любом случае, это положение не ставит под сомнение собственные биоинформатические результаты Соискателя, полученные в данной работе.

В качестве еще одного замечания, следует отметить определенную тематическую несвязность очень интересных глав функциональной геномики (ChIP-Seq, ChIP-PET и ChIA-PET) и главы, связанной с выбором проб для платформы Аффиметрикс. Логичной связи здесь нет, более того, технология микрочипов по мере удешевления секвенирования будет представлять исключительно исторический интерес.

Заключение.

Представленная диссертация является законченным исследовательским трудом, представляет собой законченное, систематическое и важное научное исследование. Полученные автором результаты достоверны, выводы и заключения обоснованы. Автореферат верно отражает основное содержание диссертации.

Работа обсуждалась на расширенном заседании семинара Лаборатории геномики и эпигеномики позвоночных Центра «Биоинженерия» РАН, с приглашением всех заинтересованных сотрудников и получила одобрение участников семинара, протокол заседания № 1 от 29 января 2014 года. Отзыв заслушан и обсужден на заседании семинара Лаборатории геномики и эпигеномики позвоночных (21.10.2014, протокол № 9).

Диссертация Орлова Юрия Львовича, на тему «Полногеномный компьютерный анализ распределения сайтов связывания транскрипционных факторов эукариот по данным иммунопреципитации хроматина и высокопроизводительного секвенирования», представленная к защите на соискание учёной степени доктора биологических наук, соответствует всем требованиям ВАК (пп. 9-14 Положения о присуждении ученых степеней, утверждённого Постановлением Правительства РФ от 24.09.2013 г. № 842), предъявляемым к докторским диссертациям, а её автор достоин присуждения искомой степени доктора биологических наук по специальности 03.01.09 - Математическая биология, биоинформатика.

10.11.2014 г.

Заведующий лабораторией
геномики и эпигеномики позвоночных
Центра «Биоинженерия» РАН, д.б.н.



Прохорчук Е.Б.