Федеральное агентство научных организаций
Федеральное государственное бюджетное учреждение науки
Институт молекулярной биологии им. В.А. Энгельгардта
Российской академии наук (ИМБ РАН)

на правах рукописи

Кулаковский Иван Владимирович

Регуляторные мотивы в геномах высших эукариот и их роль в экспрессии генов

03.01.09 – «Математическая биология, биоинформатика»

Диссертация на соискание ученой степени
доктора биологических наук

Москва – 2017

Содержание

Резюме

Исследование структуры и функции генома на основе его последовательности – одна из ключевых областей современной биоинформатики и молекулярной биологии. Настоящая работа посвящена разработке и практическому применению вычислительных методов анализа характерных коротких паттернов – мотивов – в нуклеотидных последовательностях. Методическая часть фокусируется на идентификации и поиске мотивов в современных экспериментальных данных по ДНК-белковому узнаванию, сравнении мотивов и оценке точности их вычислительного представления. Практическая часть посвящена применению разработанных методов для аннотации регуляторных последовательностей в различных задачах геномики высших эукариот. В работе представлена систематическая коллекция мотивов, описывающих участки связывания факторов транскрипции человека и мыши, и для конкретных регуляторных систем проведен анализ роли мотивов в регуляции транскрипции.

1. Предисловие

1.1. Биоинформатика как дисциплина

Использование ЭВМ во многом изменяет характер умственного труда ученого. … Небольшие группы ученых с помощью машин, совершающих более 1 млн. операций в секунду, могут выполнять трудоемкую исследовательскую работу. … В самом деле, объем знаний растет невиданно быстрыми темпами, их хранение, переработка, совершенствование и эффективное использование становится все более необходимой задачей.

И. Г. Герасимов, **Научное исследование** (Политиздат, Москва, 1972).

Удивительно, как быстро и как тесно сложные вычислительные устройства оказались интегрированы в повседневную жизнь. Вычислительная роль вычислительных машин сегодня почти забыта в глубоком подвале длинного списка экономических и социальных активностей, обеспечиваемых компьютерной инфраструктурой: от организации международных банковских платежей до личного фитнес-трекера и облачного хранилища документов. Подробная информация о всевозможных аспектах частной жизни впервые в истории человечества стала системной и структурированной, по сути, силами самих индивидов, ежечасно документирующих свою жизнь в социальных сетях и, неявно, во множестве других информационных систем, от глобальных поисковых интернет-сайтов до магазинов одежды. В этом контексте, анализ данных – содружество математики и информатики для генерации знаний на основе данных – приобрел реальное могущество, проделав путь от условно-безобидной таргетированной рекламы до массового эксперимента по манипуляции эмоциями пользователей Facebook [Kramer, Guillory, Hancock, 2014]. Массивы цифровых данных собираются, анализируются и вращают шестеренки на стыке реального и цифрового миров; а гордые в прошлом Электронные Вычислительные Машины все меньше ассоциируются с наукой, быть может за исключением арифметических монстров из суперкомпьютерного рейтинга Top500.[1] И все же, компьютерные методы, пусть и лишенные пафоса, остаются неотъемлемой частью научных исследований в разнообразных областях знаний. Более того, конкретные, обманчиво узкие тематики порождают широкий спектр вычислительных задач и привносят азарт, достаточный поддержания в боевой форме самостоятельной

[1] TOP500 Supercomputer Sites. http://www.top500.org/

научной области. Живым и в чем-то уникальным примером является биоинформатика.

Био-*информатика* как концепция была исходно сформулирована в широком смысле, охватывая различные вопросы изучения информационных процессов в биологических системах, и приобрела важную роль в эволюционной биологии (история предмета увлекательно изложена у [Hogeweg, 2011]). Анализ биологических последовательностей с помощью вычислительных методов получил признание благодаря классическим работам Маргарет Окли Дэйхоф [Hunt, 1983], в эпоху первых экспериментальных методов для прочтения последовательностей биополимеров. Экспериментальные методы подстегнули развитие биоинформатики и на текущем витке: современные высокопроизводительные экспериментальные методы требуют новых компьютерных методов для обработки результатов.

Биоинформатика невозможна без фундамента экспериментальных методов и генерируемых данных. Более жесткий вопрос, нужна ли и возможна ли самостоятельная биоинформатика? Мы смело утверждаем, что и возможна и нужна, ведь она не ограничивается *ремеслом* или *инженерной технологией* по использованию разнородных компьютерных инструментов для обработки экспериментальных результатов. Биоинформатика, благодаря тесному родству с анализом данных, является полноценной и самостоятельной дисциплиной, порождающей новое биологическое знание при грамотной постановке задач и некоторой удаче. В этой работе мы старались продемонстрировать обе стороны биоинформатики, и инструментальное инженерное дело (компьютерный анализ экспериментальных данных) и содержательную сторону биоинформатических результатов в молекулярной биологии.

1.2. Омики для регуляторной геномики

Передовая роль омиксных или омиковых данных (или просто *омиков*, -omics) и высокопроизводительных технологий, в том числе быстрых и дешевых методов прочтения геномных последовательностей уже превратилась в устоявшийся штамп научной и даже научно-популярной литературы. Скорость развития экспериментальных методов такова, что место технологий параллельного «секвенирования нового поколения» (next-generation sequencing) занимают

технологии «новейшего поколения» и процесс все ускоряется. Впрочем, невозможно отрицать массовое и чрезвычайно успешное применение высокопроизводительного секвенирования для решения широчайшего спектра задач современной молекулярной биологии [Goodwin, McPherson, McCombie, 2016]. Устойчивое удешевление стоимости прочтения одного нуклеотидного основания[2] стимулирует появление все новых вариантов экспериментальных методов с яркими перспективами как для академической науки, так и для применений в клинике [Carlson, 2012; Casey и др., 2013]. В свою очередь, рост объемов данных, появление новых и улучшение существующих экспериментальных методов требуют постоянной доработки и адаптации вычислительных средств обработки результатов. Грамотная разработка вычислительных методов требует достаточно глубокого понимания эксперимента и изучаемого объекта. Образно говоря, появление нового или заметная модификация существующего экспериментального метода (wet lab), снова превращает поле деятельности компьютерных методов (dry lab) в нетронутую целину. Есть и позитивный момент: в процессе кропотливого «повторного» решения технических задач часто появляются и самостоятельные биологические наблюдения.

Эта диссертационная работа во многом построена на результатах массового применения высокопроизводительных методов секвенирования. Благодаря новым типам экспериментальных данных, открылись новые возможности для приложения методов биоинформатики в регуляторной геномике. В то же время, простой экстенсивный рост объема прочитываемых последовательностей на два порядка увеличил масштабы вычислительных задач по анализу паттернов в последовательностях нуклеиновых кислот, задач, которые, казалось бы, успешно решены более 20 лет назад.

Прочтение последовательностей геномов множества организмов открыло возможности для «полногеномных» компьютерных исследований еще задолго до появления высокопроизводительного секвенирования. Все более полная функциональная аннотация прочитанных геномов позволяет под новым углом взглянуть на механизмы, контролирующие реализацию генетической информации. Этот вопрос особенно интересен для протяженных геномов высших эукариот.

[2] Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). http://www.genome.gov/sequencingcosts/

Действительно, кодирующие последовательности составляют лишь ограниченную долю генома (порядка 1-2% для генома человека [International Human Genome Sequencing Consortium, 2004]), а для некодирующих областей регуляторная функция является одной из ключевых [Jenks, 2013]. Данная работа посвящена анализу последовательностей, задействованных в регуляции экспрессии генов на уровне транскрипции [Riethoven, 2010].

Моду в исследованиях регуляции экспрессии задает активное развитие эпигенетики (эпигеномики) [Jaenisch, Bird, 2003], сфокусированной на внешних – по отношению к последовательности нуклеотидов – механизмах контроля экспрессии: метилировании ДНК [Jones, 2012; Smith, Meissner, 2013] и модификациях гистонов, составляющих *гистоновый код* [Burgess, 2012; Jenuwein, Allis, 2001]. В то же время структура или грамматика последовательностей регуляторных участков генома все еще остается недостаточно ясной. Фундаментальный вопрос – обладают ли транскрипционные регуляторные районы выраженным регулярным кодом по аналогии с триплетной структурой кодирующих областей – до сих пор не закрыт [Harbison и др., 2004; Istrail, De-Leon, Davidson, 2007; Rister, Desplan, 2010]. Поиск ответа привлекает и экспериментальные, и вычислительные подходы, и в большой степени опирается именно на методы анализа последовательностей.

Диссертация посвящена изучению характерных паттернов, *мотивов*, в регуляторных последовательностях геномов высших эукариот и их роли в экспрессии генов. В обзоре литературы мы в первую очередь формулируем однозначный ответ на вопрос, что такое *мотивы* в биологических последовательностях. Само слово мотив (motif) в контексте анализа последовательностей используется повсеместно, и крайне удивительно, что терминологический вопрос не решен однозначно. Затем речь идет о вычислительном представлении и применении моделей мотивов, и о классических моделях, появившихся благодаря базовым экспериментальным методам для анализа особенностей узнавания ДНК регуляторными белками – факторами транскрипции. Далее обсуждаются современные методы анализа ДНК-белкового узнавания, различные особенности высокопроизводительных методов и сопутствующие вычислительные инструменты для обработки результатов. Наконец, описываются стандартные вычислительные задачи, связанные с анализом мотивов, в том числе,

сравнение мотивов между собой и аннотация геномных вариантов в регуляторных областях.

Представленная работа преимущественно опирается на новые авторские алгоритмы и программы. В разделе «Материалы и методы» сделан акцент на математической и технической стороне разработанных подходов для анализа мотивов в нуклеотидных последовательностях (идентификация, поиск и вычислительное представление мотивов). В разделе «Результаты и обсуждение» рассматривается практическое применение анализа мотивов к различным задачам регуляторной геномики, связанным с характерными паттернами в нуклеотидных последовательностях и их ролью в регуляции экспрессии генов высших эукариот. Основное внимание уделено вопросам регуляции транскрипции, но взаимосвязь регуляции транскрипции и трансляции также затронута.

1.3. Технические замечания

Отметим несколько ключевых технических моментов по оформлению текста:

(1) таблицы приведены в тексте, рисунки расположены на отдельных страницах в конце каждого подраздела за исключением небольших врезок, иллюстрирующих конкретный абзац; (2) форматирование ссылок на литературу в тексте и записей в списке литературы приближено к стандарту ГОСТ, за исключением прямых ссылок на Интернет-ресурсы: в силу ограниченного срока жизни они не включены в основной список литературы, а приведены в виде гиперссылок в сносках; (3) в тексте используются наиболее популярные варианты русскоязычных терминов и, в тех случаях, где это показалось уместным, в скобках приведены устоявшиеся аналоги из англоязычной литературы.

В соответствии с концепцией свободного использования произведений (Россия) и концепцией добросовестного использования (США, fair use concept) в данной работе, преследующей образовательные и научные цели, используются фрагменты других работ (в т.ч. рисунки выполненные автором, соавторами или прочими исследователями) с указанием источника и авторства без запроса разрешения на использование у соответствующих правообладателей.

1.4. Список англоязычных терминов и сокращений

AUC, area under curve – площадь под кривой, часто используется в ROC-анализе (см. ниже) как численная характеристика качества классификации.

binding profile – профиль связывания, в литературе используется в качестве замены термина «мотив» в смысле паттерна схожих слов или для описания большого (например, полногеномного) набора вхождений мотива в протяженную последовательность.

ChIP-Seq, Chromatin ImmunoPrecipitation followed by massively parallel/deep Sequencing – иммунопреципитация хроматина с последующим глубоким секвенированием, основной современный метод исследования ДНК-белковых взаимодействий *in vivo* в полногеномном масштабе.

composite elements – композитные элементы – колокализованные на заданном расстоянии или перекрывающиеся сайты связывания взаимодействующих факторов транскрипции, совместно распознающих ДНК.

DIC, discrete information content – ДИС, дискретное информационное содержание – метод оценки консервативности безделеционного множественного локального выравнивания, основанный на ненормированных частотах (отсчетах) букв-нуклеотидов в колонках. KDIC, КДИС – ДИС с кульбаковским членом («кульбаковское» дискретное информационное содержание), включает член, учитывающий кульбаковское расстояние от фонового (напр. геномного) до наблюдаемого распределения нуклеотидов в колонке выравнивания.

enhancer, CRM, cis-regulatory module – энхансер, цис-регуляторный модуль гена или генов, обогащенный сайтами связывания факторов транскрипции и локализованный в некотором удалении от участка непосредственной инициации транскрипции.

GMLA, gapless multiple local alignment – безделеционное множественное локальное выравнивание (последовательностей).

homotypic clusters – гомотипические кластеры, множество близко расположенных или частично перекрывающихся сайтов связывания конкретного фактора транскрипции.

motif – мотив, характерный короткий паттерн, соответствующий схожим участкам одной или нескольких последовательностей, и/или описание совокупности этих участков в какой-либо форме.

motif core – ключевой участок (ядро, кор) мотива, наиболее консервативный между различными, но похожими словами (подпоследовательностями), использованными при выделении паттерна.

motif discovery – выявление (идентификация) паттернов в заданных последовательностях и/или построение модели паттерна.

motif family – семейство мотивов. В литературе может означать как множество мотивов-паттернов для структурного семейства факторов транскрипции, так и «множество-семейство» схожих слов, соответствующих одному конкретному мотиву.

motif finding – поиск мотива (поиск мотивом) – использование известного мотива или его модели для поиска вхождений (слов), соответствующих паттерну в заданных последовательностях.

motif hits, motif occurrences – вхождения мотива в последовательности, т.е. слова, схожие с паттерном.

NGS, next-generation sequencing – технологии секвенирования нового поколения.

PWM, position weight matrix, PSSM, position-specific scoring matrix – ПВМ, позиционно-весовая матрица – метод описания мотива в форме матрицы, в которой строки/столбцы (или наоборот) соответствуют позициям/нуклеотидам (в общем случае можно использовать любой алфавит), а значения – оценкам (предпочитаемости) конкретного нуклеотида в конкретной позиции.

ROC, receiver operating characteristic – операционная характеристика приемника – стандартный метод оценки точности классификаторов по зависимости доли ложноположительных и доли истинных положительных предсказаний.

2. Введение

2.1. Факторы транскрипции и мотивы связывания у высших эукариот

… связь между геном и внешним признаком заключается не в отдельных независимых цепях реакций, которые ведут от определенного гена к определенному признаку, но значительно сложнее.

… в осуществлении определенного наследственного признака совместно участвуют очень многие, доступные экспериментальному учету отдельные модифицирующие факторы. У нас нет основания и возможности предполагать, что при влиянии на проявление каждого отдельного признака речь идет о специфических, касающихся только данного признака «специальных генах-модификаторах», ибо мы пришли бы к бессмысленному огромному общему числу генов. Таким образом, нужно предположить, что каждый ген в разной степени должен принимать участие в целом ряде процессов развития.

Н.В. Тимофеев-Ресовский, **Общие закономерности проявления генов**. Allgemeine Erscheinungen der Gen-Manifestierung, Handbuch der Erbbiologie des Menschen. Berlin, Springer, 1940, С. 32-72 (перевод с немецкого Н.В. Глотова).

Управление активностью генов через специфические некодирующие элементы генома – базовый принцип регуляции экспрессии, продемонстрированный в классических работах Жакоба и Моно [Jacob, Monod, 1961] более полувека назад. Более того, принципиальная необходимость и возможность взаимодействия множества генов при специфической реализации генетической информации была сформулирована еще в 1940 году Н.В. Тимофеевым-Ресовским.

Процесс контролируемой реализации генетической информации является многоуровневым [Lelli, Slattery, Mann, 2012], и центральной стадией является транскрипция. Сегодня, с развитием экспериментальных методов, совместная работа разнообразных регуляторных элементов и белков-регуляторов у высших эукариот изучается в деталях как для конкретных генов-мишеней, так и в масштабе полного генома.

Активность многостадийного процесса транскрипции (от сборки преинициаторного комплекса до терминации) на каждой стадии модулируется различными белками и белковыми комплексами; важнейшей стадией является инициация транскрипции, в которой основную роль играют факторы транскрипции (иначе говоря, транскрипционные факторы).

Транскрипционные факторы традиционно делятся на два класса. Первый узкий класс составляют базальные факторы (в первую очередь белки группы TFII [Nikolov, Burley, 1997; Thomas, Chiang, 2006]), непосредственно участвующие в сборке преинициаторного белкового комплекса с РНК-полимеразой II и инициации транскрипции конкретного гена. Второй класс составляют специализированные факторы транскрипции, способные взаимодействовать и друг с другом, и с разнообразными компонентами транскрипционной машинерии, и активировать либо репрессировать транскрипцию комбинаторно, в зависимости от состава конкретного белкового комплекса. Специфичность действия факторов транскрипции на экспрессию конкретных генов реализуется через узнавание сайтов связывания – характерных участков ДНК в регуляторных сегментах [Hochheimer, Tjian, 2003; Lemon, Tjian, 2000; Taatjes, Marr, Tjian, 2004].

У высших эукариот охарактеризовано более сотни тысяч специализированных регуляторов транскрипции [Weirauch и др., 2014]. Только лишь для генома человека систематически каталогизировано более полутора тысяч белков [Vaquerizas и др., 2009; Wingender, Schoeps, Donitz, 2012]), которые способны участвовать в регуляции транскрипции напрямую либо опосредовано, связывая соответствующие ДНК-сайты в некодирующих районах. Роль факторов транскрипции не ограничивается этапом инициации и распространяется в том числе и на элонгацию транскрипции [Меркулов, Меркулова, 2014; Nechaev, Adelman, 2011], а некоторые факторы транскрипции способны участвовать в распознавании и внесении гистоновых меток [Medvedeva и др., 2015] или работать «первооткрывателями» (pioneer transcription factors), т.е. связываться с нуклеосомами, инициировать ремоделирование хроматина и привлекать другие факторы транскрипции для первичной активации экспрессии [Drouin, 2014; Soufi и др., 2015; Zaret, Carroll, 2011].

Именно координированная работа различных факторов транскрипции является базовым механизмом дифференцировки клеток и последующего поддержания клеточной идентичности [Levine, Cattoglio, Tjian, 2014; Ravasi и др., 2010]. Возникает фундаментальный вопрос: как обеспечивается время- и место-специфическая регуляция, то есть, как именно и какие именно из сотен возможных регуляторов контролируют транскрипцию конкретного гена в конкретном типе клеток в конкретный момент времени. Ответ на этот вопрос можно искать на разных

уровнях. На глобальном уровне интерес представляет экспрессия и активность самих белков-регуляторов и доступность хроматина. На промежуточном уровне можно изучать конкретные эпигенетические маркеры регуляторных районов [Jaenisch, Bird, 2003]. И наконец, на локальном геномном уровне специфическая регуляция реализуется через последовательность ДНК, которая является платформой для сборки различных комплексов факторов транскрипции, узнающих соответствующие сайты связывания [Kato и др., 2004; Wolberger, 1998]. Устоявшаяся петлевая модель [Levine, Cattoglio, Tjian, 2014; Wasserman, Sandelin, 2004] предполагает сближение удаленных регуляторных районов с проксимальными промоторами (участками генома, непосредственно окружающими регионы инициации транскрипции). Условная схема представлена на **Рисунке 1**.

Интересно, что удаленные регуляторные районы – энхансеры (enhancer [Khoury, Gruss, 1983], дословно, «усиливающий агент») – альтернативно называются цис-регуляторными модулями (cis-regulatory module, CRM [Ludwig, Patel, Kreitman, 1998]). Концепция цис-регуляции (в противовес транс-) через элементы, локализованные на той же молекуле (в цис-положении), по всей видимости была предложена Дж. Б. С. Холдейном еще в начале прошлого века [Dronamraju, 1992]. Более строго как «цис-» можно рассматривать элементы, расположенные вместе с геном в пределах одного цистрона как области генетического сцепления [Benzer, 1955; Lewis, 1951]). В реальности расстояние от энхансера до целевого промотора является вариабельной величиной, от сотен до сотен тысяч нуклеотидов [Bulger, Groudine, 2011]. Кроме того, особым случаем является трансвекция – транс-действие энхансера между гомологичными хромосомами у *Drosophila* [Müller, Schaffner, 1990]. Судя по последним данным трансвекция возможна для большого множества энхансеров [Blick и др., 2016], и это несколько путает устоявшуюся терминологию.

Многие факторы транскрипции (как минимум несколько сотен для человека, без учета неполноты существующей аннотации[3]) обладают двоякой активностью в смысле регуляции транскрипции, т.е. и функцией активатора, и функцией репрессора в зависимости от геномного контекста сайтов связывания [Stampfel и др., 2015]. При

[3] Поиск по базе данных UniProt [The UniProt Consortium, 2012], июль 2016, запрос (“transcription factor” and “human” and GO:0000122 and GO:0045944), где термины генной онтологии (GO, Gene Ontology) соответствуют активации и репрессии транскрипции, осуществляемой полимеразой II. http://uniprot.org

этом, термин *энхансер* в смысле «регуляторный участок генома» употребляется чаще, чем формально более общий вариант «цис-регуляторный модуль»; это перекликается с современным пониманием активирующей функции факторов транскрипции как основной [Hurst и др., 2014].

На уровне последовательностей регуляторных районов (энхансеров или промоторов) задача состоит в идентификации конкретных сегментов, ДНК-сайтов связывания, распознаваемых факторами транскрипции. Обычно сайты связывания представляют собой сравнительно короткие участки ДНК (10-20 пар оснований). Последовательности различных сайтов схожи для конкретного фактора транскрипции и для белков с ДНК-связывающими доменами одного семейства [Kulakovskiy и др., 2013a; Wingender, Schoeps, Donitz, 2012]. Формализованное описание сходства последовательностей сайтов связывания, т.е. наблюдаемый общий для нескольких сайтов паттерн, называется мотивом, профилем или вычислительной моделью сайта связывания. Наиболее похожий участок между сайтами связывания, т.е. наиболее выраженная и стабильная часть паттерна, часто называется ядром, кором мотива (motif core).

С точки зрения структурной биологии, ДНК-связывающие домены факторов транскрипции определяют первичную специфичность взаимодействия, то есть аффинность связывания конкретным белком конкретного участка ДНК [Luscombe и др., 2000; Rohs и др., 2010] и коровый паттерн ДНК-сайтов. Локальные особенности трехмерной структуры макромолекул и их прямые следствия, например, оптимальная ориентация водородных связей между аминокислотами белка и нуклеотидами или геометрические параметры бороздок ДНК, отражаются в предпочитаемых последовательностях сайтов связывания [Oshchepkov и др., 2004]. То есть, степень сходства различных сайтов связывания прямо (непосредственный контакт ДНК и белка) или косвенно (физические свойства локального участка ДНК) отражает предпочтения белка к распознаваемому участку ДНК [Stormo, 2013] и определяет паттерн, узнаваемый белком в регуляторных последовательностях.

История структурных исследований самой РНК-полимеразы II, осуществляющей транскрипцию белок-кодирующих генов, генов микро- и длинных некодирующих РНК, насчитывает десятки лет: от первых успешных работ [Kim,

Nikolov, Burley, 1993; Kim и др., 1993] до нобелевской премии[4] Артура Корнберга 2006 года [Bushnell и др., 2004; Cramer, Bushnell, Kornberg, 2001; Gnatt и др., 2001] и публикаций современных структур субнанометрового разрешения [Louder и др., 2016]. Структурный анализ комплексов факторов транскрипции с ДНК развивается не менее бурно, число опубликованных структур [Berman и др., 2000] составляет уже тысячи[5] и продолжает расти.

С появлением детальных аннотаций ДНК-белковых контактов [Kirsanov и др., 2013; Spirin и др., 2007] становится возможной систематическая реконструкция ДНК-сайтов связывания факторов транскрипции на основе трехмерных структур ДНК-белковых комплексов [Alamanova, Stegmaier, Kel, 2010; Morozov, Siggia, 2007; Xu и др., 2013]. Структурный анализ позволяет выявить контакты между элементами белковой структуры и конкретным фрагментом ДНК и, в ряде случаев, оценить аффинность белка к различным олигонуклеотидам. Тем не менее, структурный подход ограничен практическими затратами на получение всевозможных комплексов ДНК-белок с различными фрагментами ДНК. Сложности представляет и большое разнообразие ДНК-белковых контактов даже внутри одного семейства ДНК-связывающих доменов [Zanegina и др., 2016]. В то же время, анализ последовательностей ДНК масштабируется все лучше и, благодаря стандартизации и удешевлению экспериментов, появляется все больше данных о распределении сайтов связывания в геномах [O'Malley и др., 2016; Yan и др., 2013] и среди искуственных олигонуклеотидов [Berger и др., 2006]. В этой работе акцент сделан именно на вычислительном анализе нуклеотидных последовательностей сайтов связывания. Опираясь на экспериментальные данные, можно конструировать и применять модели паттернов для полногеномного поиска сайтов связывания *in silico* [Daily и др., 2011; Xie, Rigor, Baldi, 2009]. В свою очередь, знание локализации и аффинности сайтов связывания [Stormo, 2000] имеет массу приложений в функциональных исследованиях. Это и изучение регуляторного потенциала геномных вариантов в некодирующих областях [Macintyre и др., 2010; Ponomarenko и др., 2003; Vorontsov и др., 2015], и предсказание локализации регуляторных районов [Frith, Li, Weng, 2003;

[4] The Nobel Prize in Chemistry 2006. Nobelprize.org. Nobel Media AB 2014. http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2006/

[5] Поиск по базе данных структур PDB (Protein Data Bank), июль 2016, запрос "transcription factor". http://pdb.org

Girgis, Ovcharenko, 2012] и анализ структуры и «грамматики» регуляторных районов [Shelest, Albrecht, Shelest, 2010; Yokoyama, Ohler, Wray, 2009], и определение генов-мишеней регуляторов с последующей реконструкцией генных сетей для исследования экспрессии генов методами системной биологии [Liao и др., 2003; Liu и др., 2015].

Представленная работа основывается на синергии высокопроизводительных экспериментов по определению последовательностей сайтов связывания факторов транскрипции высших эукариот и вычислительных методов идентификации и представления паттернов в нуклеотидных последовательностях, в первую очередь, мотивов связывания факторов транскрипции. Практическая апробация разработанных компьютерных методов сделана в рамках конкретных исследований регуляции экспрессии генов у высших эукариот.

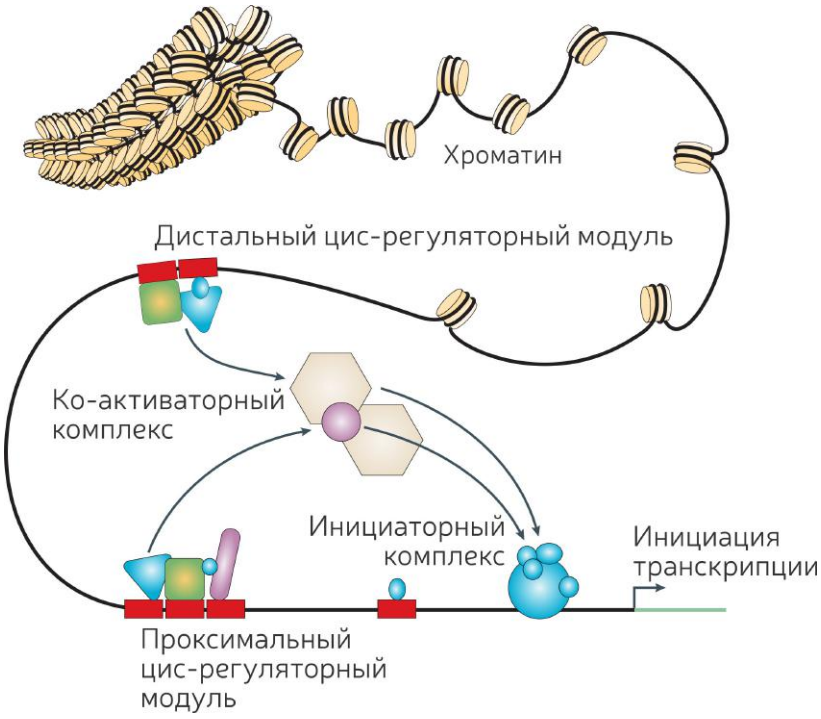


Рисунок 1. Регуляция инициации транскрипции у высших эукариот.
Инициация транскрипции управляется факторами транскрипции через стабилизацию инициациторного комплекса. Комбинации факторов транскрипции связывают цис-регуляторные модули (энхансеры). Взаимодействие проксимальных и дистальных энхансеров и промотора осуществляется через петли в трехмерной укладке хроматина. Схема адаптирована из классического обзора [Wasserman, Sandelin, 2004].

2.2. Актуальность темы

Многоуровневая регуляция экспрессии генов является ключом к управляемой реализации генетической информации, которая определяет координированное развитие разнообразных типов клеток высших эукариот. Базовым звеном в регуляции экспрессии является регуляция транскрипции генов, которая в большой степени определяется некодирующими районами генома, связывающими белковые факторы. Благодаря появлению доступных методов для массового прочтения последовательностей ДНК, стремительно растет объем прямых данных по ДНК-белковому узнаванию как *in vivo*, так и *in vitro*. Компьютерный анализ характерных ДНК-паттернов, мотивов, распознаваемых факторами транскрипции, потенциально позволяет изучать структуру регуляторных районов с однонуклеотидным разрешением. Однако, классические вычислительные инструменты для анализа мотивов не справляются с возрастающими объемами данных и не учитывают специфику современных экспериментальных подходов. При этом, область применения анализа мотивов не ограничивается конкретными случаями ДНК-белкового узнавания или отдельными регуляторными районами конкретных генов. С накоплением экспериментальных данных становится возможным систематический анализ для выявления глобальных закономерностей в колокализации мотивов и других функциональных элементов генома и изучения регуляции транскрипции в геномном масштабе на уровне последовательности: от анализа грамматики регуляторных районов до функциональной аннотации геномных вариантов. В свою очередь, эта информация является важным компонентом для реконструкции генных сетей и индивидуальной геномики. Совокупно, это обуславливает высокую актуальность разработки и применения новых компьютерных методов для анализа специфических нуклеотидных паттернов, задействованных в регуляции экспрессии генов.

2.3. Цель и задачи работы

Цель работы: выявление, характеристика и систематизация мотивов в некодирующих районах геномов высших эукариот для решения задач регуляторной геномики путем вычислительного анализа данных, полученных современными высокопроизводительными экспериментальными методами.

Задачи работы:

(1) разработка биоинформатических методов для идентификации, поиска и сравнения паттернов-мотивов в нуклеотидных последовательностях;

(2) создание систематической коллекции мотивов связывания факторов транскрипции мыши и человека на основе опубликованных экспериментальных данных, включая

результаты современных высокопроизводительных экспериментов по иммунопреципитации хроматина;

(3) практическая апробация разработанных методов в конкретных задачах регуляторной геномики:

а. выявление особенностей колокализации мотивов ключевых факторов плюрипотентности OCT4/SOX2/NANOG;

б. установление связи кластеризации сайтов связывания фактора Spi1 с экспрессией генов в мышиной модели эритролейкемии;

в. определение давления отбора на соматические мутации в сайтах связывания различных транскрипционных факторов в геномах раковых клеток;

г. поиск взаимосвязи регуляции транскрипции и трансляции на примере сигнального каскада mTOR;

д. изучение колокализации сайтов связывания факторов транскрипции и CpG-светофоров;

е. систематическая идентификация мотивов в ткань-специфичных промоторах, полногеномно определенных для мыши и человека с помощью технологии кэп-анализа экспрессии генов.

2.4. Научная новизна, теоретическое значение и научно-практическая ценность работы

В ходе работы был разработан комплекс **новых биоинформатических методов** для анализа мотивов в нуклеотидных последовательностях. Путем интеграции и кросс-валидации данных различных экспериментальных источников, построена **новая**, наиболее полная коллекция мотивов ДНК-белкового узнавания для факторов транскрипции мыши и человека. Созданные в ходе работы методы нашли широкое практическое применение и позволили установить ряд **новых фактов** о локализации мотивов в регуляторных районах генов и их роли в экспрессии генов. В том числе, **впервые** на основе данных по иммунопреципитации хроматина систематически идентифицированы тройственные композитные элементы сайтов связывания факторов транскрипции OCT4/SOX2/NANOG; установлено избегание ключевых позиций мотивов сайтов связывания относительно CpG-светофоров; выявлено действие отрицательного отбора на соматические мутации, возникающие в сайтах связывания ряда семейств факторов транскрипции в геномах раковых клеток; показана контрастная роль кластеров сайтов связывания белка Spi1 в регуляции экспрессии генов при эритролейкемии.

Предложенные вычислительные методы успешно использованы для анализа мотивов в регуляции экспрессии генов мыши и человека. Возможная сфера применения

разработанных методов значительно шире: это и геномы других эукариот, например, растений, для которых появляется массовая экспериментальная информация о регуляции, и геномы прокариот. Наличие методической базы и наиболее полной и точной коллекции мотивов открывает новые возможности как для решения конкретных задач (аннотации конкретных некодирующих вариантов или конкретных промоторов отдельных генов), так и для глобального анализа регуляторных районов. Мотивы могут быть спроецированы на структуры ДНК-белковых комплексов для совместного изучения различных типов контактов ДНК-белок и локальных особенностей олигонуклеотидов, отраженных в их последовательностях. Сходство ДНК-связывающих доменов у факторов транскрипции внутри структурного семейства позволяет использовать представленные в работе мотивы для анализа регуляции транскрипции и у менее изученных видов живых организмов.

Теоретическое значение и научно-практическая ценность диссертации подтверждаются активным цитированием ключевых статей[6], грантовой поддержкой работ (первый конкурс грантов для молодых биологов фонда «Династия» Дмитрия Зимина, 2012; ряд проектов, поддержанных Российским научным фондом и Российским фондом фундаментальных исследований, в т.ч. в роли руководителя) и наградами научного сообщества: премия Европейской Академии (2016), Медаль «Феномен жизни» памяти В.И. Корогодина (2015), почетная грамота Российской Академии Наук (2015).

Все представленные в работе вычислительные методы документированы и опубликованы в сети Интернет как программы с открытым исходным кодом. Это обеспечивает свободный доступ к методической части работы для широкого исследовательского сообщества, и позволяет ее практическое использование в научной и образовательной деятельности.

2.5. Апробация и публикации по теме работы

Список публикаций по теме диссертации включает **21** статью в рецензируемых международных журналах, **2** приглашенные главы-обзора [Kulakovskiy, Makeev, 2013; Kulakovskiy, Makeev, 2014], **2** статьи в российских журналах, **2** статьи в рецензируемых сборниках конференций. Автором сделано **22** доклада, включая устные и приглашенные, на конференциях в России и зарубежом, среди которых «Биология – наука 21 века» (Пущино, 2017), BGRS (Новосибирск, 2016, 2012, 2010), SocBiN Bioinformatics (Москва, 2016), MCCMB (Москва, 2015, 2013, 2011), ISMB/ECCB (Дублин, 2015; Берлин, 2013; Вена, 2011),

[6] Ivan Kulakovskiy – Google Scholar Citations.
https://scholar.google.ru/citations?user=0f5hVB4AAAAJ&hl=ru

«Современные проблемы генетики, радиобиологии, радиоэкологии и эволюции» (Санкт-Петербург, 2015), BIOSTEC BIOINFORMATICS (Барселона, 2013), POSTGENOME (Казань, 2012), ECCB (Базель, 2012; Гент, 2010), “Albany 2011: The 17th conversation” (Олбани, США), ESF FG&D (Дрезден, 2010). Полный список авторских публикаций и докладов по теме диссертации приведен в соответствующем разделе.

Материалы диссертации активно используются в образовательном процессе. Автором прочитаны приглашенные лекции по анализу мотивов и ChIP-Seq данных в ходе образовательных курсов: «Анализ данных в биоинформатике и практические приложения» (школа в рамках конференции SocBiN Bioinformatics, Москва, 2016), «Биоинформатика высокопроизводительного секвенирования» (Школа биоинформатики, Москва, 2016), «Анализ данных высокопроизводительного секвенирования» (ФББ МГУ, 2015), «Анализ ОМИКСных данных в медицине» (Сколково, 2015), на Летней школе биоинформатики (Москва, 2016), на Школе молекулярной и теоретической биологии (проект Фонда Дмитрия Зимина «Династия», Пущино, 2012-2015).

2.6. Личный вклад автора

В методических работах [Kulakovskiy и др., 2010; Kulakovskiy и др., 2011; Kulakovskiy и др., 2013b; Kulakovskiy и др., 2013c] автором диссертации лично выполнена разработка, программная реализация алгоритмов, тестирование и статистический анализ. В методических работах [Vorontsov и др., 2015; Vorontsov, Kulakovskiy, Makeev, 2013] автор диссертации принимал прямое участие в разработке алгоритма, дизайне и документировании программной реализации и тестировании.

В работе [Kulakovskiy и др., 2013a] автором диссертации предложен подход к организации коллекции мотивов и сопутствующих исходных данных, выполнена массовая вычислительная идентификация мотивов, сравнительное тестирование и, частично, экспертное курирование результатов. В работе [Kulakovskiy и др., 2016] автором диссертации проведена идентификация мотивов, разработан подход для систематического сравнительного тестирования, проведено экспертное курирование полученных мотивов.

В работах [Медведева и др., 2010; Afanasyeva и др., 2017; Kozlov и др., 2014; Kozlov и др., 2015; Levitsky и др., 2014; Maksimenko и др., 2015; Medvedeva и др., 2010; Medvedeva и др., 2014; Ridinger-Saison и др., 2012; Schwartz и др., 2016; Schwartz и др., 2017] автором диссертации выполнен вычислительный анализ мотивов с помощью инструментов, созданных в рамках диссертации (в т.ч. идентификация мотивов и поиск вхождений). В работе [Eliseeva и др., 2013] автором диссертации поставлена задача и координирован процесс исследований. В работе [Forrest и др., 2014], опубликованной консорциумом

FANTOM, автором диссертации проведена идентификация мотивов в промоторах, активных в различных типах клеток, предложена и частично реализована процедура интеграции результатов идентификации мотивов, полученных различными программными инструментами. В работе [Medvedeva и др., 2015] автор принимал участие в разработке структуры базы данных и интеграции информации о факторах транскрипции. Для работы [Vorontsov и др., 2016] автором диссертации предложена общая схема исследования и дизайн вычислительного эксперимента.

Автор диссертации принимал непосредственное участие и в биологической интерпретации результатов упомянутых выше работ, и в написании и редактировании текстов публикаций. В **7** статьях по теме диссертации автор выступает в качестве первого автора, и в **9** в качестве автора, ответственного за переписку (corresponding author).

3. Обзор литературы

Регуляция транскрипции у высших эукариот и сопутствующая роль регуляторных мотивов переплетены с различными аспектами молекулярной биологии и бионформатики. В этом обзоре основное внимание уделено факторам траскрипции и ДНК-мотивам связывания. В стороне намеренно оставлены такие темы, как эпигенетическая регуляция [Jaenisch, Bird, 2003] и регуляция с помощью микроРНК [Sevignani и др., 2006]. При обсуждении факторов транскрипции опущен ряд интересных подробностей: не освещены особенности трехмерных структур ДНК-связывающих доменов [Stegmaier, Kel, Wingender, 2004; Wingender, 2013], механизмы стохастического поиска белками сайтов связывания в геноме [Normanno и др., 2015], специфические особенности локальной структуры ДНК в районе сайтов связывания [Rohs и др., 2009; Yang и др., 2014], функциональное взаимодействие факторов транскрипции и длинных некодирующих РНК [Ng и др., 2013], экспериментальное картирование и вычислительное предсказание энхансеров [Frith, Li, Weng, 2003; Shlyueva, Stampfel, Stark, 2014; Suryamohan, Halfon, 2015], а также использование информации о сайтах связывания при реконструкции регуляторных сетей [Santra, 2014; Verfaillie и др., 2015].

Мы не можем не признавать важность и масштаб этих тем, раскрывающих роль факторов транскрипции и мотивов связывания в фундаментальных вопросах генной регуляции, но надеемся, что в обзоре литературы удалось сфокусироваться на моментах, наиболее близких к основному содержанию и методическим достижениям диссертационной работы: экспериментальным особенностям и вычислительной обработке данных высокопроизводительных экспериментов о последовательностях, узнаваемых факторами транскрипции в ДНК, а также вычислительных методах представления, идентификации и сравнения мотивов.

3.1. Мотивы и структура регуляторных последовательностей

3.1.1. Терминологический вопрос

В контексте вычислительного анализа нуклеотидных последовательностей часто возникает необходимость специфически называть специфические короткие участки нуклеотидных последовательностей (т.е. «слова текста» ДНК или РНК), например, сайты связывания факторов транскрипции в энхансере исследуемого гена. Методы анализа и описания коротких паттернов в нуклеотидных последовательностях развиваются уже более 30 лет [Stormo и др., 1982], в том числе и для представления участков связывания факторов транскрипции [Berg, Hippel von, 1987; Stormo, Schneider, Gold, 1986]. Удивительно, но массово употребляемый термин «мотив» до сих пор лишен однозначной трактовки. Представим себе множество последовательностей участков связывания конкретного фактора в регуляторных районах генома. Некоторые авторы [Sinha, Tompa, 2002] используют термин «мотив» в смысле «паттерн», т.е. описание набора схожих последовательностей как единой сущности. Тогда конкретные сайты связывания и их последовательности называются вхождениями мотива (motif hits, motif occurrences). Другой вариант использования слова «мотив» предполагает, что это конкретный участок последовательности ДНК, соответствующий конкретному сайту связывания [Wang, Yu, Zhang, 2005]. В таком случае можно было бы говорить о «модели» ДНК-белкового узнавания на уровне последовательности ДНК, описывающей множество сайтов связывания, где «мотивом» является каждый конкретный сайт. Однако, вместо этого некоторые авторы для обозначения общего паттерна используют «семейство мотивов» (motif family) [Monteiro и др., 2008]. Чтобы окончательно запутать читателей, в некоторых работах [D'haeseleer, 2006; Xie, Rigor, Baldi, 2009] в пределах одного текста мотивом называется и паттерн (модель) и его вхождения (последовательности конкретных сайтов связывания). Неудивительно, что ряд авторов вообще избегает слова «мотив» и стремится предложить альтернативные термины, например «профиль связывания» (binding profile) для обозначения общего паттерна сайтов [Wasserman, Sandelin, 2004].

С терминами следующего уровня, идентификацией мотивов (motif discovery, выявление общего паттерна) и поиском мотивами (или мотивным поиском, т.е.

поиском вхождений мотива, motif finding) также существует путаница: «*ab initio* discovery» соседствует с «*ab initio* finder» и оба варианта означают идентификацию мотива – выявление паттерна [Brown и др., 2013]. Тем не менее, чаще всего «*de novo* motif discovery» [Kuttippurathu и др., 2011] и «*ab initio* motif discovery» [Machanick, Bailey, 2011] используются для обозначения выявления паттерна и построения его вычислительной модели, а поиск вхождений известного паттерна называется «motif finding» [Kulakovskiy, Favorov, Makeev, 2009].

В этой работе мы придерживаемся следующей терминологической схемы: *мотивом* называется общий паттерн для функционально-близкого набора последовательностей (в том числе, соответствующих схожим сайтам связывания конкретного фактора транскрипции) и его вычислительное представление; *вхождением мотива* называется конкретный участок ДНК, схожий с паттерном; *идентификацией мотива* называется определение сходства между последовательностями и построение вычислительной модели паттерна.

3.1.2. Промоторы и энхансеры эукариот

Некодирующие сегменты генома могут нести регуляторную информацию, используемую на различных уровнях экспрессии генов: от управления организацией контактов между различными сегментами ДНК (инсуляторы) до транскрипции (энхансеры и промоторы) и трансляции (5’ и 3’ нетранслируемые районы мРНК). Наша работа посвящена центральному уровню регуляции: контролю инициации транскрипции, осуществляемой РНК-полимеразой II, транскрибирующей белок-кодирующие гены, гены микроРНК и длинных некодирующих РНК.

Традиционная схема регуляции говорит о выпетливании (запетливании, looping) ДНК и формировании контактов между удаленным энхансером (цис-регуляторным районом) и промотором (участком ДНК, где непосредственно происходит инициация транскрипции гена). Задача выделения функциональных петель «промотор-энхансер», регулирующих активность транскрипции в сложной структуре хромосомных контактов, пока в общем случае не решена; как и не ясно, какая или какие из множества моделей образования петель функциональны [Mora и др., 2015]. Однако, для конкретных генов и типов клеток комплексы «промотор-

энхансер-факторы транскрипции» удается успешно реконструировать, см. **Рисунок 2**.

3.1.2.1. Эукариотические промоторы

Десятилетия активных исследований все еще не дали полной ясности в описании структуры регионов непосредственной инициации транскрипции у эукариот, т.н. коровых промоторов. Исторически, наиболее известным промоторным мотивом является TATA-бокс [Lifton и др., 1978; Mathis, Chambon, 1981], сайт посадки TATA-связывающего белка (TBP, TATA-box binding protein) – компонента базального комплекса TFIID. В ближайшей окрестности был позднее найден ряд других мотивов, как сопутствующих TATA-боксу (BRE, TFIIB recognition element), так и достаточных для самостоятельного запуска инициации (Inr, Initiator [Smale и др., 1998; Yang и др., 2007]). Помимо общих элементов были найдены и специфичные к таксону, например для позвоночных (DCE, downstream core element) и мух (MTE, motif ten element; DPE, downstream promoter element) [Lenhard, Sandelin, Carninci, 2012; Smale, Kadonaga, 2003]. Несмотря на консервативность в локализации относительно сайта инициации транскрипции, элементы коровых промоторов присутствуют в последовательности не всегда и вхождения мотивов могут заметно отличаться от консенсуса [Butler, Kadonaga, 2002]. Это верно даже для канонического TATA-бокса, который первоначально считался основным и неотъемлемым элементом промоторов, но уверенность в его определяющей роли стремительно падала с ростом объема экспериментальных данных об активности участков инициации транскрипции [Trinklein и др., 2003]. Современные оценки отводят TATA-промоторам менее 30% от общего множества [Yang и др., 2007]. Неясна судьба и других коровых элементов: ряд паттернов (GC-бокс, CCAAT-бокс) в литературе часто продолжают относить к коровым промоторам [Wu и др., 2006], хотя уже известно, что они не относятся к базальной машинерии, а являются мотивами сайтов связывания крупных семейств специфических факторов транскрипции.

Сегодня можно уверенно говорить о выраженных классах промоторов млекопитающих, основываясь на их функциональности и точности, т.е. ширине региона, где происходит инициация транскрипции [Lenhard, Sandelin, Carninci, 2012]: (1) ткань-специфичные гены, экспрессируемые в дифференцированных тканях взрослого организма, обычно имеют узкие TATA-промоторы; (2) гены «домашнего

хозяйства», активно экспрессируемые в различных тканях и на различных стадиях развития, имеют широкие CG-богатые не-TATA промоторы; (3) гены, кодирующие рибосомные белки и факторы трансляции, имеют узкие TCT-промоторы, GC-богатые, обогащенные TATA-боксами и стабильно высоко экспрессируемые в большинстве типов клеток. У *Drosophila* можно выделить похожие функциональные группы, но для них характерны другие комбинации свойств.

С точки зрения анализа последовательностей интересно, что локальный нуклеотидный контекст и вхождения мотивов (в т.ч. коровых элементов) могут быть успешно использованы для предсказания позиционирования стартов инициации транскрипции [Frith и др., 2008; Megraw и др., 2009], причем информация о сайтах связывания факторов транскрипции вносит заметный вклад в точность предсказаний.

Говоря о высших эукариотах, важно еще раз явно отметить, что понимание старта транскрипции гена как конкретной единичной позиции генома является условно допустимым только для узких TATA- и TCT-промоторов. GC-богатый промотор часто обеспечивает инициацию в широкой области с множеством активных субрегионов, покрывающих в сумме сегменты генома длиной более сотни нуклеотидов [Sandelin и др., 2007]. В ходе международного проекта FANTOM (Functional annotation of the mammalian genome) использование метода кэп-анализа экспрессии генов (cap analysis of gene expression [Kawaji и др., 2011]) и технологии секвенирования одной молекулы Helicos (без ПЦР-амплификации) позволило построить детальный «промотором» (promoterome), карту и количественную оценку активности областей инициации транскрипции в геномах человека и мыши [Forrest и др., 2014]. В сотнях клеточных линий и первичных клеток удалось достоверно идентифицировать более сотни тысяч отдельных участков инициации транскрипции, среди которых как множественные альтернативные старты транскрипции известных белок-кодирующих генов, так и промоторы некодирующих РНК [Hon и др., 2017]. Управление экспрессией в таком масштабе требует координированной многоуровневой регуляции, задействующей и глобальные эпигенетические механизмы, и сайты связывания факторов транскрипции.

3.1.2.2. Транскрипционная активность энхансеров

Совершенствование высокопроизводительного секвенирования и сопутствующих молекулярно-биологических методов привело к обнаружению повсеместной транскрипционной активности генома, не объяснимой исходя из карты белок-кодирующих генов. Возник естественный вопрос, является ли эта фоновая активность «побочным шумом» от нормальной работы РНК-полимеразы [Ponjavic, Ponting, Lunter, 2007; Struhl, 2007] или же несет функциональную нагрузку [Kapranov и др., 2007]. Среди функциональных аспектов, с одной стороны, активно изучается транскрипция и разнообразие длинных некодирующих РНК (нкРНК) [Hon и др., 2017; Mercer, Dinger, Mattick, 2009]. С другой стороны – энхансерные РНК (эРНК), и транскрипционная активность энхансеров [Andersson и др., 2014].

В классическом понимании, энхансеры это участки ДНК, способные дистанционно влиять на транскрипцию гена, но относительно близко расположенные на той же хромосоме. Факт транскрипции энхансеров долгое время ускользал от достоверного наблюдения, поскольку такие транскрипты быстро деградируют [Wyers и др., 2005]. Сегодня, благодаря глубокому секвенированию РНК удалось уверенно выделить два класса транскрибируемых энхансеров [Natoli, Andrau, 2012], см. **Рисунок 3**. Первый класс сбалансированно продуцирует «двунаправленные» эРНК (2d-eRNA), по которым удается не только локализовать сам энхансер в геноме, но и оценить его регуляторную активность (которая скоррелирована с активностью транскрипции эРНК). Второй класс более загадочен: однонаправленная эРНК (1d-eRNA) очень похожа на длинную некодирующую РНК. Фактически, возник новый класс задач: определить функциональность самого энхансера, как регуляторного элемента, а не транскрибируемой с него нкРНК [Paralkar и др., 2016] и, обратно, выделить независимую роль транскрибируемой эРНК на фоне регуляторной активности энхансера [Hsieh и др., 2014]. Уже высказано предположение об участии эРНК непосредственно в формировании петель «промотор-энхансер», но имеющиеся свидетельства противоречивы [Daniel, Nagy, Nagy, 2014]. Таким образом, вопрос об основной роли эРНК можно считать открытым. Еще один фундаментальный вопрос, можно ли систематически выявить функциональные взаимосвязи между энхансерами и генами-мишенями с помощью вычислительных методов и существующих

экспериментальных данных, также еще не получил однозначного ответа [Daniel, Nagy, Nagy, 2014; Mora и др., 2015].

Активная транскрипция и наличие сайтов связывания факторов транскрипции, как ключевых регуляторных элементов, позволяют говорить о функциональной близости промоторов и энхансеров [Kim, Shiekhattar, 2015], параллельно идентифицировать их по транскрипционной активности, используя единые экспериментальные данные [Andersson и др., 2014], и применять схожие биоинформатические методы для анализа мотивов.

3.1.3. Грамматика регуляторных областей

С точки зрения анализа последовательностей интерес представляет локальная структура регуляторных областей, а именно, взаимная локализация сайтов связывания факторов транскрипции и связанные с этим функциональные свойства энхансеров. В представлении моделей многофакторной регуляции существует две крайности: «билборд» (billboard) и «энхансеосома» [Arnosti, Kulkarni, 2005; Panne, 2008], см. **Рисунок 4**. Они отличаются степенью структурированности комплекса факторов транскрипции и сложностью организации регуляторного района ДНК. «Коллектив» факторов транскрипции представляет собой промежуточный (между билбордом и энхансеосомой) вариант нестрогой организации энхансера. Возможны и крайности, например, высокопроизводительное тестирование активности тысяч энхансеров, узнаваемых опухолевым супрессором p53, показало достаточность единственного сайта связывания [Verfaillie и др., 2016]. Судя по всему, различные биологические системы требуют разнообразных типов энхансеров, что позволяет в широких пределах устанавливать паттерны экспрессии генов от эмбрионального развития до стрессового ответа.

В структурированных энхансерах, в том числе в предельном случае «энхансеосомы», связывание комплекса факторов транскрипции с ДНК устанавливает достаточно жесткие ограничения на относительную локализацию соответствующих сайтов связывания. Это позволяет выделять наборы, например, пары сайтов, где соответствующие вхождения мотивов локализованы на характерном расстоянии друг от друга, то есть составляют так называемые композитные элементы (composite elements [Kel-Margoulis и др., 2002]). Предпочтительные расстояния

между сайтами, образующими композитные элементы, имеют различную степень выраженности в зависимости от конкретных факторов транскрипции и регуляторной системы [Kulakovskiy и др., 2011; Shelest, Albrecht, Shelest, 2010; Yokoyama, Ohler, Wray, 2009].

Параллельное наблюдение состоит в том, что регуляторные районы ДНК часто характеризуются увеличенной плотностью сайтов связывания для конкретного фактора транскрипции, то есть обнаруживаются группы похожих сайтов связывания, называемые гомотипическими кластерами (homotypic clusters), которые исходно были обнаружены у дрожжей [Wagner, 1997], а затем активно изучались в энхансерах генов раннего развития *Drosophila* [Lifanov и др., 2003]. Позднее гомотипическая кластеризация сайтов связывания была обнаружена и в геноме человека [Gotea и др., 2010], в том числе в промоторных областях генов. Степень кластеризации сайтов связывания зависит от конкретного фактора транскрипции [Murakami, Kojima, Sakaki, 2004], но интересно, что сайты связывания в кластерах у позвоночных могут быть более эволюционно консервативны, чем одиночные сайты [Gotea и др., 2010]. Сегодня известна масса примеров, когда гомотипические кластеры в промоторах играют принципиальную роль в установлении паттернов экспрессии генов [Giorgetti и др., 2010; Lang, Juan, 2010; Lettice и др., 2012; Loo van и др., 2012; Schindler, Sherwood, 2011]. Однако, как и в случае с энхансерами, известны случаи, когда ключевую роль играет единственный выраженный сайт связывания [Whitfield и др., 2012].

С появлением экспериментальных методов для массового параллельного тестирования потенциала энхансеров *in vivo* [Inoue, Ahituv, 2015; Kheradpour и др., 2013] можно выяснить степени свободы для положения и аффинности вхождений мотивов как базовых элементов регуляторного кода. Интересно, что функциональные энхансеры допускают баланс «грамматики» и «орфографии», т.е. относительного положения сайтов связывания и их аффинности [Farley и др., 2016]. Анализ мотивов сегодня уже предоставляет инструменты для детального изучения структуры энхансеров и промоторов. В смелой перспективе – целевая модификация и рациональный дизайн регуляторных последовательностей.

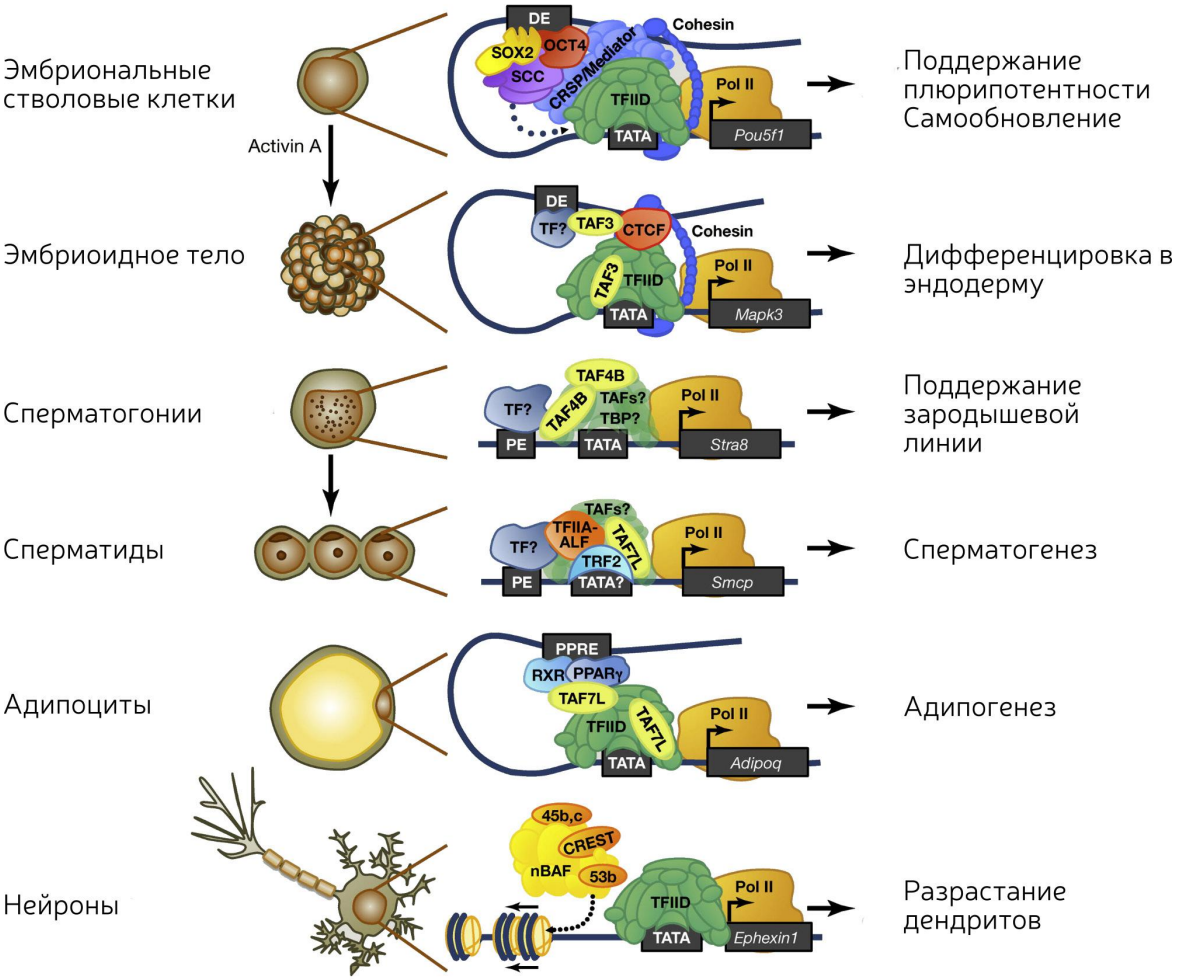


Рисунок 2. Разнообразие машинерии, обеспечивающей реализацию программ транскрипции, специфичных для различных типов клеток, на примере конкретных генов.

Панели сверху-вниз: поддержание плюрипотентности эмбриональных стволовых клеток и дифференцировка в эндодерму; поддержание идентичности зародышевой линии и сперматогенез; адипогенез; разрастание дендритов нейронов. На схемах в качестве отдельных элементов показаны различные типы белков: специфичные факторы транскрипции (напр. SOX2 и OCT4); когезин, стабилизирующий петли энхансер-промотор; РНК-полимераза II; базальные факторы транскрипции (TFIID), комплексы ремоделирования хроматина (nBAF), проксимальные и дистальные энхансеры (PE, DE). Схема адаптирована из обзора [Levine, Cattoglio, Tjian, 2014].

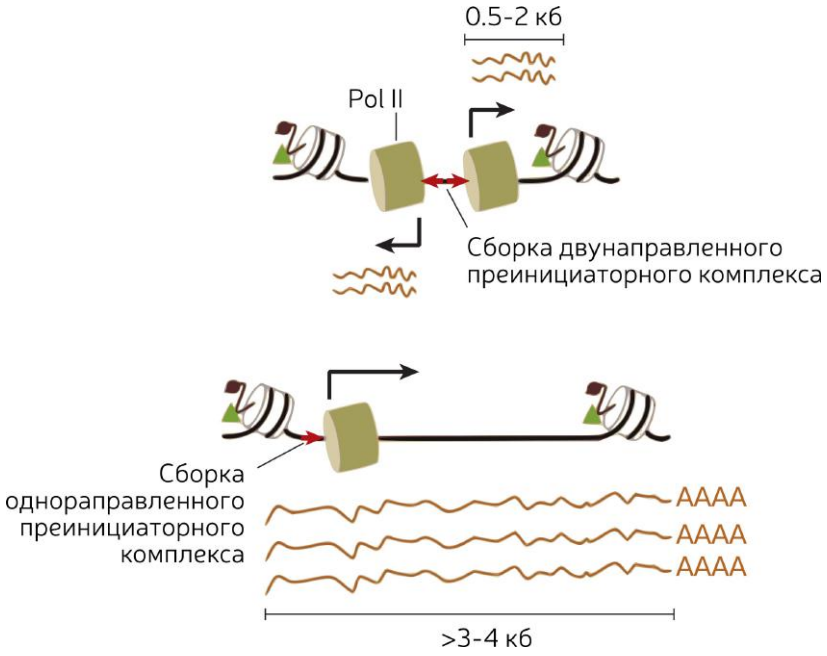
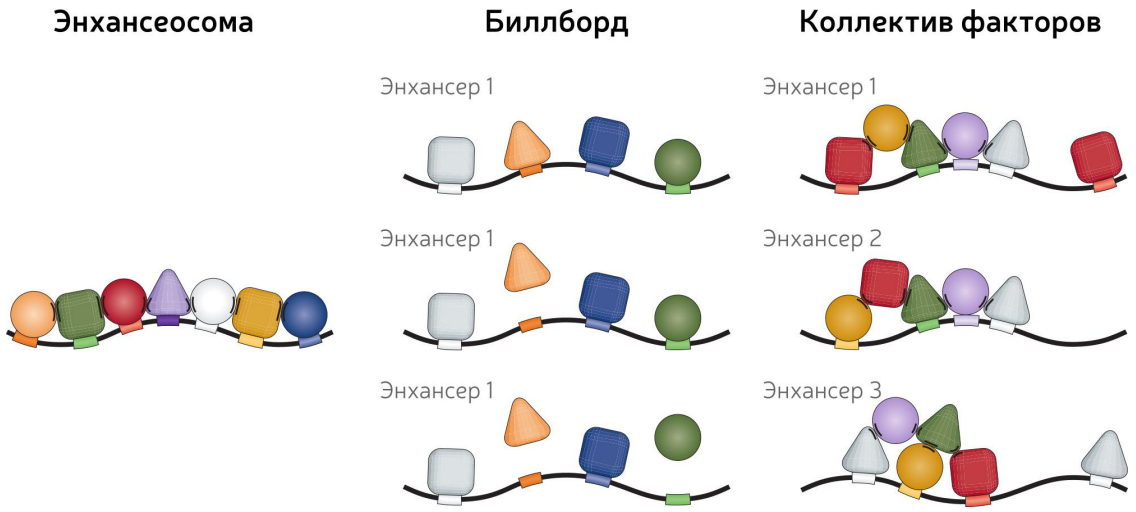


Рисунок 3. Характерные типы некодирующих РНК, транскрибируемых с энхансерных районов (эРНК).

(верхняя панель) Двунаправленные, относительно короткие, неполиаденилированные эРНК (2d-eRNAs). *(нижняя панель)* Однонаправленные, длинные, полиаденилированные эРНК (1d-eRNAs). Направленность транскрипции может определяться как самим преинициаторным комплексом, так и специфическими факторами транскрипции. Схема адаптирована из обзора [Natoli, Andrau, 2012].



	Энхансеосома	**Билборд**	**Коллектив факторов**
Взаимодействие белок-ДНК	Сильное кооперативное связывание ДНК, выполняющей роль платформы	Кооперативное и аддитивное связывание	Кооперативное связывание ДНК, и ДНК и белковый комплекс могут выполнять роль платформы
Белковый интерфейс	Фиксирован, определен комплексом ДНК-белки	Изменчивый, возможны варианты	Изменчивый, возможны варианты
Мотивы	Фиксированная грамматика: сайты связывания всех белков должны присутствовать на определенных местах	Фиксированный состав мотивов, гибкая грамматика	Гибкий состав мотивов (различные факторы транскрипции могут напрямую связывать ДНК), гибкая грамматика
Активность	Единая (требует присутствия и активности всех факторов транскрипции, составляющих энхансеосому)	Активность не требует присутствия всех факторов транскрипции	Коллективная активность: требуется присутствие нескольких факторов транскрипции, но неясно, обязательно ли присутствие всех

Рисунок 4. Существующие модели энхансера.

(левая панель) Энхансеосома. Соответствует ситуации, в которой все факторы, связывающие энхансер, необходимы для кооперативного связывания и активации. Фиксированная грамматика – структурированный набор сайтов связывания работает как платформа для сборки белкового интерфейса высокого порядка, обеспечивающего регуляцию транскрипции [Merika, Thanos, 2001]. (*средняя панель*) Билборд. Подразумевает гибкое позиционирование сайтов связывания – ограниченного поднабора сайтов достаточно для транскрипционной активности [Arnosti, Kulkarni, 2005]. *(правая панель)* Коллектив факторов транскрипции. Один и тот же набор факторов способен связывать различные энхансеры в различных конфигурациях. В случае, когда энхансер содержит сайты связывания не для всех факторов, недостающие регуляторы могут привлекаться через белок-белковые взаимодействия [Junion и др., 2012].

3.2. Вычислительное представление и практический анализ мотивов

Essentially, all models are wrong, but some are useful.
По существу, все модели ошибочны, но некоторые – полезны.

[Box, Draper, 1987]

3.2.1. Мотив как множество вырожденных подстрок

У высших эукариот типичные сайты связывания сравнительно коротки, а их мотивы «вырождены», т.е. допускают неточные совпадения. Это контрастирует с четко определенными сайтами, характерными, например, для эндонуклеаз рестрикции. Способность факторов транскрипции узнавать сайты с различной аффинностью и конкурировать за связывание похожих участков ДНК делает возможной гибкую количественную регуляцию экспрессии. Вырожденность и сравнительно малая длина сайтов связывания (10-20 пар оснований) ограничивает применение стандартных методов для поиска гомологичных участков последовательностей [Cliften и др., 2001] на основе множественного локального выравнивания [Thompson, Gibson, Higgins, 2002]. Тем не менее, задача идентификации мотива, т.е. паттерна, описывающего локальную область сходства между последовательностями сайтов связывания, часто формулируется именно как построение множественного безделеционного локального выравнивания [Das, Dai, 2007]. Следует сразу сказать, что для ДНК-мотивов сайтов связывания факторов транскрипции обычно совместно рассматривают мотив и его обратно-комплементарное отражение (ориентацию), поскольку большинство факторов транскрипции осуществляют регуляцию через связывание именно двуцепочечной ДНК.

При рассмотрении множественного локального выравнивания достаточно естественным выглядит базовое предположение, что высоко консервативные позиции выравнивания (с предпочтительным содержанием конкретных нуклеотидов) соответствуют позициям сайтов связывания, которые являются более важными для ДНК-белкового узнавания. Фиксация конкретных наиболее консервативных нуклеотидов дает простейшие модели представления мотива как паттерна на четырехбуквенном нуклеотидном алфавите {A,C,G,T}: консенсусные

последовательности в нотации IUPAC[7] с допустимыми заменами [Day, McMorris, 1992] и регулярные выражения [Myers, Miller, 1989]. Такие модели легко редуцируются до ограниченного списка допустимых подстрок (слов), могут быть построены по считанным экспериментально определенным сайтам связывания, но только грубо, в бинарной форме «да/нет», описывают эффект от возможных замен нуклеотидов. Тем не менее, для слабо изученных или чрезвычайно хорошо выраженных сайтов связывания консенсусные модели до сих пор используются: например, консенсусы представлены в популярных коллекциях мотивов связывания для факторов транскрипции растений [Hehl, Bülow, 2014; Yilmaz и др., 2011].

Естественное развитие консенсусных моделей состоит в том, чтобы явно учесть количественную информацию о нахождении различных нуклеотидов в конкретной позиции выравнивания различных сайтов связывания для одного фактора транскрипции. Замечательно, что эта простая идея оказалось достаточной для создания базовой модели мотива – позиционно-весовой матрицы, которая не просто описывает множество вырожденных подстрок-сайтов, но и позволяет количественно оценить энергию ДНК-белкового взаимодействия [Berg, Hippel von, 1987; Stormo, Schneider, Gold, 1986]. Следующий раздел посвящен описанию именно этой модели, которая уже 30 лет наиболее массово и успешно используется для представления мотивов ДНК-белкового узнавания.

3.2.1.1. Позиционно-весовые матрицы

Позиционно-весовые матрицы (ПВМ; или просто позиционные матрицы, PWM, position weight matrix, и они же PSSM, position-specific scoring matrix) представляют собой простой способ представления мотива сайтов связывания длины m нуклеотидов в форме матрицы $4\times m$, где строки соответствуют нуклеотидам {A,C,G,T}, а столбцы – позициям сайта связывания (часто используется и транспонированная запись где строки – позиции сайта, а столбцы – нуклеотиды). Числа в матрице – веса – соответствуют предпочтению или избеганию конкретного нуклеотида в конкретной позиции. Каждому олигонуклеотиду-слову длины m ПВМ сопоставляет оценку (PWM score, скор) – вещественное число, которое получается

[7] Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences. http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html

как произведение или сумма весов в соответствующих ячейках матрицы. Использовать позиционно-весовые матрицы для описания мотива связывания впервые предложил Гэри Стормо [Stormo и др., 1982] для распознавания РНК-сайтов инициации трансляции у прокариот (*E. coli*). Для подбора значений весов в этой работе использовалась одна из первых искусственных нейросетей – перцептрон [Rosenblatt, 1958], а полученные ПВМ заметно превзошли точность доступных на тот момент консенсусных моделей.

В дальнейшем были предложены альтернативные методы подбора весов [Harr, Haggstrom, Gustafsson, 1983] и удалось показать, что при правильном подборе весов ПВМ-оценка сайта в промоторе коррелирует с активностью этого промотора у *E. coli* [Mulligan и др., 1984]. Ключевые работы [Berg, Hippel von, 1987; Berg, Hippel von, 1988] обеспечили ПВМ биофизическим фундаментом: с помощью методов статистической механики авторы показали, что оценка ПВМ пропорциональна аффинности сайтов.

ПВМ основаны на идее независимости соседних нуклеотидов, как с точки зрения вероятности их нахождения в функциональных сайтах, так и с точки зрения их вклада в энергию взаимодействия белка и ДНК.

Наиболее популярный сегодня способ построения ПВМ использует логарифмические отношения вероятностей (log-odds) [Stormo, 2000]. Пусть имеется множественное безделеционное локальное выравнивание сайтов связывания. По нему можно напрямую построить матрицу наблюдаемых частот нуклеотидов (отсчетов, position count matrix, PCM). Весовая матрица получается из матрицы нормализованных частот (вероятностей) с помощью логарифмического преобразования: $S_{\alpha,j} = \log\left(x_{\alpha,j}/N\right)$ где $x_{\alpha,j}$ – число наблюдений нуклеотида α в позиции j, а N – полное число последовательностей в выравниваниии. В этом случае оценка тестируемого олигонуклеотида (оценка сайта) определяется как сумма весов соответствующих нуклеотидов на соответствующих позициях: $\sum_{j=1}^{m} S_{\mathrm{w}[j],j}$ где w – тестируемый олигонуклеотид длины m.

По различным причинам наблюдаемые в ограниченной выборке сайтов частоты конкретных нуклеотидов могут оказаться нулевыми, кроме того, вероятности случайного наблюдения нуклеотидов неодинаковы в зависимости от GC-состава генома или конкретных изучаемых последовательностей. Таким образом,

в формулу добавляются (1) псевдоотсчеты (псевдонаблюдения) для компенсации нулевых значений и (2) нормализация наблюдаемых частот на ожидаемые из фонового (например, геномного) распределения.

В наших работах используется преобразование как у [Lifanov и др., 2003]: $S_{\alpha,j} = log\left(\frac{x_{\alpha,j}+cq_\alpha}{(N+c)q_\alpha}\right)$, где q_α принимаются равным геномным или равномерным частотам нуклеотидов, $\sum_{\alpha\epsilon\{A,C,G,T\}} q_\alpha = 1$; а c – псевдоотсчет. Нетрудно заметить, что формула для $S_{\alpha,j}$ дает нейтральную (нулевую) оценку для случая, когда наблюдаемая частота нуклеотида $x_{\alpha,j}$ в точности соответствует ожидаемой из геномного распределения Nq_α.

Геномы высших эукариот не являются однородными по нуклеотидному составу, и частоты нуклеотидов могут отличаться даже в регуляторных областях, обладающих схожей функцией (вспомним о различных классах промоторов). Поэтому, на практике удобно использовать равномерные фоновые частоты нуклеотидов, и делать поправку на локальный нуклеотидный состав не при определении весов ПВМ, а при оценке статистической значимости конкретных вхождений мотивов.

Построение ПВМ схематично проиллюстрировано на **Рисунке 5**. Вопрос о выборе псевдоотсчета не является тривиальным [Nishida, Frith, Nakai, 2009]. В своей работе мы используем псевдоотсчеты $\ln N$, которые масштабируются с размером выборки (что удобно при совместном использования мотивов, построенных по ограниченным догеномным и масштабным постгеномным наборам данных).

Несмотря на естественные ограничения весовых матриц, проистекающие из фиксированной длины мотива и предположения о независимости соседних позиций, ПВМ все еще доминируют в анализе мотивов. Базовый инструментарий для использования ПВМ, например идентификация и визуализация мотивов, уже достаточно развит (об этом пойдет речь в следующих разделах). В то же время в некоторых направлениях методы для ПВМ продолжают активно разрабатываться, например, появляются альтернативные схемы для оценки энергии связывания [Zhao, Granas, Stormo, 2009] и способы масштабирования существующих ПВМ для сравнения оценок энергии связывания между различными мотивами или факторами транскрипции [Ma и др., 2015].

3.2.1.2. Информационное содержание и визуализация мотивов в форме лого-диаграмм

Основа весовых матриц – множественное выравнивание сайтов связывания и матрица частот. Параллельно с использованием весов как логарифмов отношений возник и альтернативный взгляд на выравнивание с точки зрения информационного содержания (или энтропии Шеннона). Для конкретной j-й позиции сайта связывания информационное содержание равно $I_j = 2 + \sum_{\alpha\epsilon\{A,C,G,T\}} f_{\alpha,j} \log_2 f_{\alpha,j}$, обозначения соответствуют введенным ранее, а $f_{\alpha,j} = x_{\alpha,j}/N$ является нормализованной частотой нуклеотида, т.е. оценкой вероятности встретить конкретный нуклеотид в j-й позиции сайта связывания. Наиболее важные колонки выравнивания являются наиболее консервативными, т.е. в пределе содержат только один нуклеотид из возможных четырех и имеют максимальное информационное содержание в два бита [Schneider и др., 1986].

В предположении о независимости соседних колонок выравнивания (т.е. соседних позиций сайта связывания) информационное содержание полного выравнивания равно сумме информационных содержаний всех колонок. С учетом возможных неравномерных фоновых частот нуклеотидов q_α (например, вызванных неравномерным GC-составом изучаемого генома) можно использовать «относительную энтропию», т.е. расстояние Кульбака-Лейбера от распределения фоновых частот [Stormo, 1990]: $I_j = \sum_{\alpha\epsilon\{A,C,G,T\}} f_{\alpha,j} \log_2 f_{\alpha,j}/q_\alpha$.

В качестве прямой системы скоринга сайтов связывания информационное содержание не прижилось, но стало стандартным способом масштабирования высоты колонок при визуализации выравниваний в форме лого-диаграмм (sequence logos [Schneider, Stephens, 1990]): по горизонтальной оси откладываются позиции, а по вертикальной оси информационное содержание колонки выравнивания (отображаемая высота колонки соответствует информационному содержанию, а высоты нуклеотидов соответствуют их относительным частотам, см. пример лого на **Рисунке 5**).

В последнее время лого-диаграммы подвергаются критике, в связи с характерными отличиями в восприятии информации о важности конкретных нуклеотидов у экспертов и начинающих исследователей [Kok, Oon, Lee, 2014]. Предлагаются альтернативные варианты отрисовки мотивов, в том числе для

визуального сравнения мотивов между собой. Тем не менее, лого-диаграммы остаются общепринятым и наиболее популярным методом визуализации мотивов и коротких локальных выравниваний.

В нашей работе мы многократно используем визуализацию мотивов в форме лого, но вместо классического информационного содержания пользуемся дискретным аналогом (дискретное информационное содержание, ДИС, discrete information content [Kulakovskiy, Favorov, Makeev, 2009]). ДИС универсально подходит для выравниваний любого числа сайтов и, в случае малых выборок, не требует специальных поправок для оценок вероятностей, поскольку оперирует исходными частотам (отсчетами): $\text{ДИС}_j = \sum_{\alpha\epsilon\{A,C,G,T\}} \frac{1}{N}\left(\ln\left(x_{\alpha,j}!\right) - \ln(N!)\right)$.

По определению, максимальное дискретное информационное содержание равно нулю (соответствует наибольшей колонке лого), за наименьшую колонку лого (с нулевой высотой) принимается «виртуальная» колонка, в которой все нуклеотиды встречаются с одинаковой частотой. Включение в дискретное инф. содержание кульбаковского члена и связь с классическим инф. содержанием подробнее обсуждается в разделе, посвященном описанию программы ChIPMunk в разделе «Материалы и методы».

3.2.1.3. Переход к расширенным моделям мотивов

Практический успех весовых матриц не отменяет ограничений, проистекающих из гипотезы о независимом аддитивном вкладе нуклеотидов в энергию связывания [Benos, Bulyk, Stormo, 2002]. ДНК-белковые взаимодействия по своей природе являются структурными, описана масса случаев, когда единственный аминокислотный остаток контактирует сразу с несколькими нуклеотидными основаниями [Luscombe, Laskowski, Thornton, 2001], например, в соседних позициях сайта связывания. Интересно, что далекие скореллированные контакты встречаются заметно реже, но также возможны [Jacobson и др., 1997]. Роль в ДНК-белковом взаимодействии играют и локальные особенности конформации ДНК, и стэкинг-взаимодействия [Rohs и др., 2009; Rohs и др., 2010]. Не исключена серьезная деформация ДНК в месте связывания белка, например, как происходит при связывании TBP с TATA-боксом [Kim и др., 1993]. Анализ наборов известных сайтов связывания позволяет выявить скоррелированные позиции [Tomovic, Oakeley, 2007],

но значимость таких наблюдений долгое время была неочевидна в силу ограниченных выборок из нескольких десятков последовательностей. По сути, именно скромный объем экспериментальных данных долгое время ограничивал применение и распространение расширенных моделей, учитывающих физико-химические свойства ДНК [Oshchepkov и др., 2004] и удаленные корреляции [Levitskii и др., 2006].

В то же время, систематический анализ новых массовых экспериментов вызвал нешуточную дискуссию [Morris, Bulyk, Hughes, 2011; Zhao, Stormo, 2011]: достаточно ли простой ПВМ и гипотезы о независимости соседних позиций для корректного описания особенностей мотивов для большинства белков? Действительно ли вычислительные методы уже достигли потолка и способны построить оптимальную ПВМ, а следующим шагом в улучшении распознавания сайтов связывания может быть только построение более сложных моделей? Это было совсем неочевидно, ведь перевыравнивание известных сайтов связывания [Keilwagen и др., 2009; Kulakovskiy, Favorov, Makeev, 2009] или дополнительное включение в выравнивание сайтов, предсказанных в геномных последовательностях [Morozov, Ioshikhes, 2013] позволяло значительно улучшить существующие ПВМ; как улучшала точность представления сайтов и тщательная оптимизация длины мотива [Levitsky и др., 2014; Morozov, Ioshikhes, 2013].

Тем не менее, современные экспериментальные данные уже не позволяют полностью игнорировать основное ограничение ПВМ – аддитивность/независимость вкладов соседних нуклеотидов. Построение расширенных моделей мотивов с использованием современных данных ChIP-Seq обсуждается ниже в отдельном разделе.

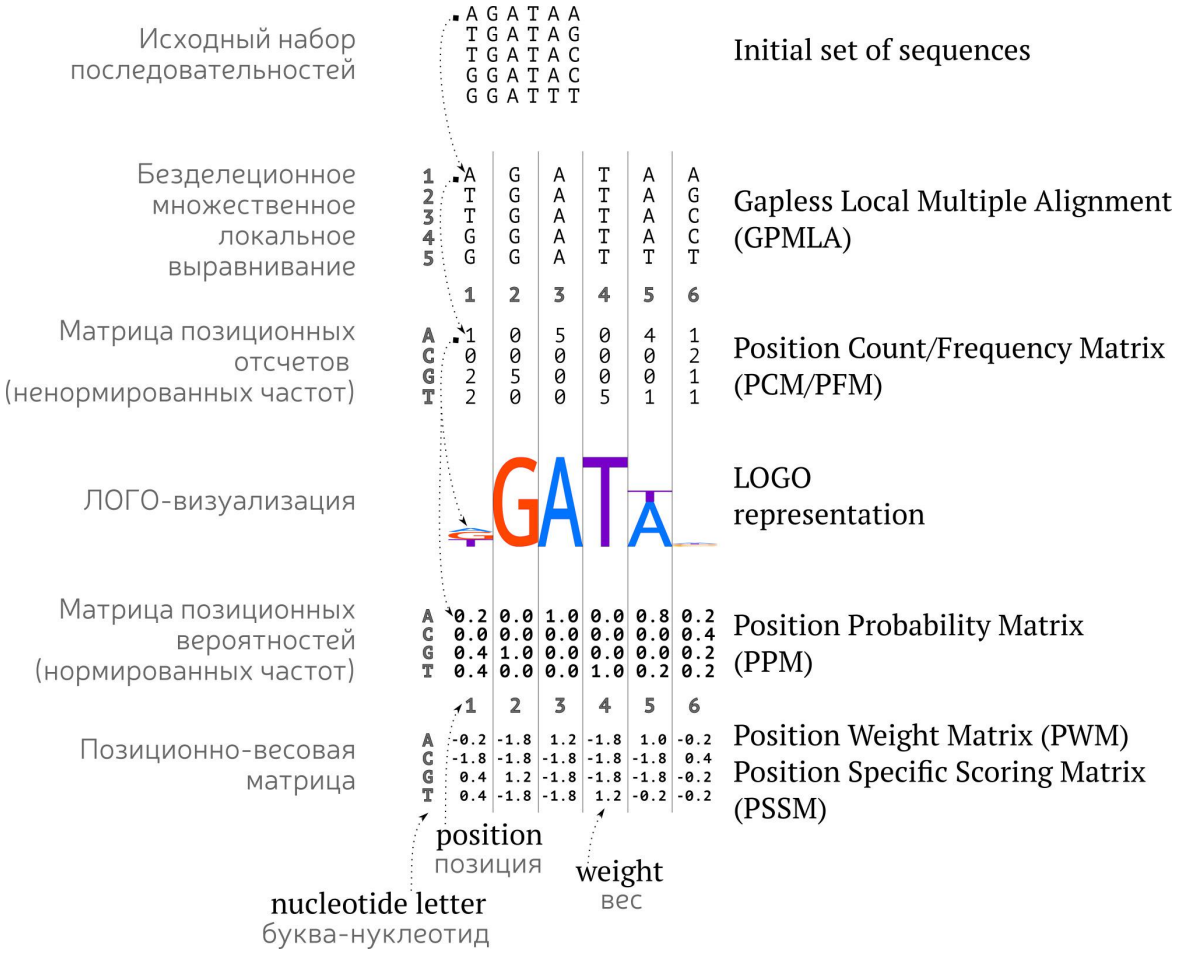


Рисунок 5. Схема построения позиционно-весовых матриц на основе безделеционного множественного локального выравнивания.

Рисунок адаптирован из обзора [Kulakovskiy, Makeev, 2013].

3.2.2. Стандартные методы идентификации мотивов

Для прикладного математика, пришедшего в биоинформатику, задача выделения мотива-паттерна как набора похожих подстрок в тексте, на первый взгляд, кажется легко формализуемой и чрезвычайно привлекательной. В работе [Das, Dai, 2007] обсуждалось уже более 150 алгоритмов и программ для идентификации мотивов. Трудно сказать, сколько всего программ опубликовано к сегодняшнему дню, и какие из них поддерживаются и находятся в работоспособном состоянии. Простота формализации задачи по идентификации мотива обманчива. Априори трудно предсказать даже то, как вхождения мотива распределены по протяженным последовательностям. Относительно универсальный подход соответствует построению безделеционного локального выравнивания последовательностей с вхождениями мотива и отбрасыванию последовательностей, не содержащих достоверных вхождений (ни одного либо одно вхождение мотива в последовательность, zero or one occurrence per sequence); увы, с ростом длины последовательностей эта постановка утрачивает смысл и необходимо учитывать множественные вхождения мотивов.

Даже в упрощенном виде – как построение выравнивания – идентификация мотива *de novo* может рассматриваться с разных сторон. С одной стороны, это задача выявления релевантного паттерна в выборке последовательностей среди множества паразитных сигналов. С другой стороны, это задача построения модели, пригодной для дальнейшего использования в качестве классификатора олигонуклеотидов по их аффинности к белку, т.е. пригодной для распознавания или предсказания сайтов связывания. Задача выявления релевантного паттерна и задача построения предсказывающей модели, безусловно, связаны, и часто решаются совместно; но полезно понимать, что их оптимальное решение не всегда возможно в рамках одного вычислительного подхода. По всей видимости, историческое смешение этих задач связано с ранними ограничениями объема экспериментальных данных: это были или считанные прокариотические сайты связывания, определенные «поштучно», либо промоторы корегулируемых генов [Favorov и др., 2005].

Задача быстрого выявления релевантного паттерна хорошо решается «словарными» методами с помощью различных техник вычислительного поиска повторяющихся подстрок в последовательностях с последующей оценкой

статистической значимости. Это и простой перебор коротких олигонуклеотидов [Helden van, Andre, Collado-Vides, 1998], и анализ встречаемости IUPAC-консенсусов [Sinha, Tompa, 2003], и использование регулярных выражений (Brazma et al., 1998). С ростом объема данных стали массово использоваться суффиксные деревья [Eskin, Pevzner, 2002; Vanet и др., 2000; Pavesi и др., 2004] и методы оптимизации на графах [Fratkin и др., 2006]. Вычислительная эффективность для словарных методов обычно не является бутылочным горлышком, но возникает проблема определения искомого паттерна среди множества альтернатив. Для различения «истинного» мотива связывания изучаемого белка и мотивов белков-кофакторов было предложено оценивать статистическую значимость наблюдаемого паттерна не только по сравнению с некоторой статистической фоновой моделью «геномного шума», но и относительно контрольной выборки реальных геномных последовательностей. Дифференциальная идентификация мотива (discriminative motif discovery) на сегодня является стандартом для наиболее успешных программ по быстрой массовой идентификации мотивов в больших наборах данных [Grau и др., 2013; Heinz и др., 2010; Thomas-Chollier и др., 2012]. В этом случае «словарная» часть алгоритма может служить для быстрого поиска перепредставленных подстрок, а наиболее удачные мотивы-кандидаты затем используются в качестве затравок для построения более точных моделей, например, стандартных позиционно-весовых матриц.

Это возвращает к вопросу о детальной модели мотива, пригодной для предсказания сайтов связывания в неаннотированных последовательностях. Параллельное направление в идентификации мотивов сразу использовало именно вероятностные, «количественные» модели, в первую очередь, – весовые матрицы. Одну из первых практических реализаций получил «жадный» алгоритм [Hertz, Hartzell G W, Stormo, 1990] с использованием информационного содержания как критерия оптимальности выравнивания. Впоследствии наибольшее распространение получили методы построения ПВМ на основе метода максимального правдоподобия (expectation-maximization, EM [Leung, Chin, 2006; Sinha, Blanchette, Tompa, 2004; Lawrence, Reilly, 1990; Bailey, 2002]) и сэмплирования по Гиббсу (gibbs sampling [Thompson и др., 2007; Thompson, Rouchka, Lawrence, 2003; Lawrence и др., 1993; Favorov и др., 2005]). Оба метода итеративно перестраивают ПВМ и соответствующее выравнивание, но в случае максимизации правдоподобия шаги

оптимизации детерминированы, а для гиббсовского подхода определяются случайным сэмплированием.

Любопытно, что в наиболее известном тестировании классических методов идентификации мотивов [Tompa и др., 2005] победил словарный алгоритм [Pavesi и др., 2004]. Однако постепенно он уступил популярность программе MEME (Multiple EM for Motif Elicitation [Bailey, 2002; Bailey, Elkan, 1994]), использующей ПВМ и максимизацию правдоподобия. Несмотря на средние результаты в сравнительных тестах MEME сохраняет популярность уже многие годы благодаря многолетнему последовательному развитию и авторской поддержке, в то числе в смысле удобства пользовательского интерфейса. Сегодня MEME Suite[8] [Bailey и др., 2009; Bailey и др., 2015] предоставляет широкий инструментарий по анализу мотивов и является, помимо RSAT [Thomas-Chollier и др., 2008], одним из немногих стабильных инструментов в биоинформатике анализа мотивов. Тем не менее, ряд специальных задач лучше решается более узкоспециализированными программами, в особенности это касается анализа мотивов в современных массовых данных.

3.2.3. Коллекции известных мотивов связывания факторов транскрипции

За несколько десятилетий исследования ДНК-белкового узнавания скопилась масса прочитанных последовательностей сайтов связывания и построенных мотивов для различных факторов транскрипции. Попытки систематизации этих знаний предпринимались различными группами [Kolchanov и др., 2002; Heinemeyer и др., 1998]. Для высших эукариот основным источником информации, курируемым вручную на основании разрозненных публикаций, долгое время была коммерческая база данных TRANSFAC [Heinemeyer и др., 1998; Matys и др., 2006; Wingender и др., 2000]. С ростом объема данных, TRANSFAC стал предлагать все больше похожих моделей мотивов для одного фактора транскрипции, что на практике усложняло интерпретацию результатов. Этой проблемы удалось избежать в другой коммерческой разработке, Genomatix MatBase[9]. Коммерческие коллекции мотивов долгое время лидировали благодаря инвестициям в анализ и систематизацию

[8] The MEME Suite. http://meme-suite.org/
[9] В составе Genomatix Software Suite.
http://www.genomatix.de/solutions/genomatix-software-suite.html

опубликованных экспериментальных данных. Бесплатные аналоги покрывали заметно меньшее число факторов транскрипции, но постепенно приобретали популярность: например, многовидовая база JASPAR [Sandelin и др., 2004] продолжает развиваться [Mathelier и др., 2015a] и активно используется.

По-настоящему востребованными открытые базы данных по мотивам и факторам транскрипции стали с появлением высокопроизводительных данных, что позволило заметно расширить спектр мотивов по структурным семействам факторов и улучшить качество распознавания сайтов связывания. Судя по результатам относительно свежих тестов [Dabrowski и др., 2015], сегодня публичные базы данных лидируют и количественно (по ширине спектра покрытых факторов транскрипции), и качественно (по точности мотивов).

Открытых коллекций мотивов, описывающих сайты связывания факторов транскрипции высших эукариот, существует множество, кратко отметим ключевые:

JASPAR[10] [Mathelier и др., 2014; Mathelier и др., 2015a] – множество видов эукариот (включая растения), прямые экспериментальные данные для различных видов;

FlyFactorSurvey[11] [Zhu и др., 2011] – плодовая мушка *D. melanogaster*, в первую очередь данные бактериальной одногибридной системы;

UniProbe[12] [Newburger, Bulyk, 2009; Hume и др., 2015] – представлены мотивы связывания факторов транскрипции разных эукариотических организмов (в основном данные для факторов транскрипции мыши и дрожжей) – результаты анализа белок-связывающих микрочипов;

SwissRegulon[13] [Pachkov и др., 2013] – мотивы для факторов транскрипции мыши и человека, построенные с учетом эволюционной консервативности сайтов связывания;

HOCOMOCO[14] [Kulakovskiy и др., 2013a; Kulakovskiy и др., 2016] – факторы транскрипции мыши и человека, интеграция различных экспериментальных

[10] The JASPAR database. http://jaspar.genereg.net
[11] Fly Factor Survey. http://mccb.umassmed.edu/ffs/
[12] UniProbe Database. http://the_brain.bwh.harvard.edu/uniprobe/
[13] SwissRegulon. http://www.swissregulon.unibas.ch
[14] HOCOMOCO COmprehensive MOtif COllection. http://hocomoco.autosome.ru

источников с фокусом на результатах иммунопреципитации хроматина с глубоким секвенированием (ChIP-Seq);

CIS-BP[15] [Weirauch и др., 2014] – наиболее полное покрытие различных видов и классов факторов транскрипции путем назначения мотивов по гомологии ДНК-связывающих доменов.

Хороший унифицированный (т.е. пригодный для автоматизированного биоинформатического использования) срез коллекций мотивов для различных организмов представлен как часть RSAT[16] [Thomas-Chollier и др., 2008] – в наборе инструментов по анализу регуляторных последовательностей.

Данные из коллекций JASPAR, SwissRegulon и TRANSFAC напрямую использованы в этой диссертационной работе, а создание коллекции HOCOMOCO является одним из основных результатов.

3.2.4. Практический анализ мотивов

Регуляторная часть генома эукариот устроена значительно сложнее, чем у прокариот: последовательности промоторов и энхансеров одновременно содержат множественные сайты связывания различных факторов транскрипции, сайты связывания реже обладают характерной структурой (например, типа прямой повтор или палиндром [Favorov и др., 2005]), регуляторные районы обладают достаточно ограниченной эволюционной консервативностью и могут располагаться на значительном удалении от целевых генов-мишеней [Odom и др., 2007; Schmidt и др., 2010]. На заре анализа регуляторных мотивов у эукариот *in silico* существовал миф о неспособности ПВМ сколько-нибудь точно предсказывать сайты связывания в связи с завышенной долей ложноположительных предсказаний в протяженных нуклеотидных последовательностях [Wasserman, Sandelin, 2004]. Отчасти это было вызвано недостаточно точными моделями мотивов, построенными по малым выборкам сайтов связывания; но, тем не менее, предсказанные *in silico* сайты успешно связывали белок *in vitro* [Tronche и др., 1997]. Сегодня уже понятно, что ограничения ПВМ как модели не являются критическими, и глобальное связывание

[15] CIS-BP, the online library of transcription factors and their DNA binding motifs. http://cisbp.ccbr.utoronto.ca

[16] Index of /rsat/motif_databases. http://floresta.eead.csic.es/rsat/motif_databases/

белков с ДНК *in vivo* в значительной степени зависит от укладки ДНК – доступности хроматина и белок-белковых взаимодействий между самими факторами транскрипции. С распространением прямых экспериментальных данных о положении нуклеосом и структуре хроматина [Meyer, Liu, 2014] задача полногеномного предсказания сайтов связывания на основании одной лишь только последовательности ДНК потеряла свою актуальность *per se*, но интерес к детальной аннотации регуляторных районов безусловно сохраняется.

На этой почве формулируется базовая задача анализа мотивов: при наличии готовой модели провести поиск вхождений мотива в заданной нуклеотидной последовательности. В отличие от идентификации мотива *de novo*, эта проблема хорошо и однозначно формализуется. Быстрые алгоритмы для поиска вхождений мотивов в форме ПВМ развивались давно [Beckstette и др., 2006], но наибольший вклад в эту область внесла группа Эско Укконена с серией работ от быстрого одновременного картирования набора ПВМ [Korhonen и др., 2009] до поиска вхождений мотивов с помощью расширенных моделей [Giaquinta, Grabowski, Ukkonen, 2013; Korhonen и др., 2016] и моделей с разделителем (спейсером, spacer) [Giaquinta и др., 2014]. На практике для поиска с помощью ПВМ часто используют программу FIMO в составе MEME Suite [Grant, Bailey, Noble, 2011] и инструменты из пакета RSAT [Medina-Rivera и др., 2015; Thomas-Chollier и др., 2008]. Авторская программа SPRy-SARUS для поиска вхождений, созданная в ходе диссертационной работы, кратко представлена ниже в разделе «Материалы и методы».

3.2.4.1. Статистическая значимость вхождений мотивов

Шкала оценки ПВМ зависит от многих параметров: длины мотива, обучающей выборки сайтов связывания, способа преобразования частот в веса. Тот факт, что ПВМ-оценка возрастает с ростом аффинности связывания, может быть использован в задачах моделирования экспрессии генов на основании кооперативного или конкурентного связывания факторами транскрипции соответствующих сайтов [Kozlov и др., 2014; Kozlov и др., 2015]. Однако в большинстве случаев явная единая шкала оценок выглядит удобнее. Одним из способов получения такой шкалы является перевод оценок в статистические значимости.

Рассмотрим ПВМ длины m и соответствующий словарь, т.е. все возможные олигонуклеотиды – слова длины m над четырехбуквенным нуклеотидным алфавитом. Каждому слову ПВМ сопоставляет число – оценку. Для предсказания сайтов связывания необходимо выбрать пороговое значение оценки t, которое будет соответствовать положительному предсказанию (слово является потенциальным сайтом связывания). Для конкретного t существует конкретная часть словаря, которая состоит из слов с оценками не хуже t. В простейшем случае, именно эта доля словаря определяет *P*-значение мотива для порога t. Если представить себе плотность распределения всех возможных оценок ПВМ, то *P*-значение соответствует площади под правым хвостом распределения, т.е. вероятности случайно выбрать из словаря слово с оценкой не хуже пороговой (см. условную схему на **Рисунке 6**). Эту идею можно обобщить на случай «неравновесных» слов, считая, что разные слова выбираются из словаря с разной вероятностью. Иначе говоря, *P*-значение это вероятность пронаблюдать слово с оценкой не хуже t в случайном месте случайной последовательности, сгенерированной согласно заданной фоновой модели. С формализацией можно познакомиться ниже, в разделе «Материалы и методы».

Поскольку оценка ПВМ является суммой независимых случайных величин (оценок нуклеотидов), то велик соблазн считать распределение значений оценки нормальным и прямолинейно выбирать пороги и *P*-значения по средней оценке и стандартному отклонению [Kulakovskiy, Makeev, 2009]. На самом деле, распределение оценок ПВМ является дискретным и может заметно отличаться от нормального именно в конечном правом хвосте, т.е. в области наиболее сильных оценок, которой принадлежат наиболее интересные сайты связывания. Для получения точного *P*-значения в работе [Touzet, Varré, 2007] было предложено использовать динамическое программирование, и в последние годы этот метод стал общепринятым. Мы реализовали этот алгоритм в [Vorontsov и др., 2015], в том числе, для вычисления *P*-значений динуклеотидных весовых матриц (см. соотв. подраздел «Материалов и методов»).

Приятное свойство *P*-значений – единая вероятностная шкала для любых видов моделей. Неприятное свойство – эффективная шкала является узкой для коротких или слабо выраженных мотивов. Это легко проиллюстрировать на примере: пусть мотив задает один конкретный олигонуклеотид. Для олигонуклеотида длины 2

P-значение на равновероятном словаре: $1/4^2 = 0.0625$. Для олигонуклеотида длины 8: $1/4^8 \sim 10^{-5}$. Фактически, для фиксированного *P*-значения более длинные мотивы всегда имеют большее разнообразие слов. Одновременно, для низких *P*-значений невозможно пронаблюдать ни одного вхождения достаточно коротких мотивов. То есть, если рассматривать единственное наилучшее вхождение мотива в изучаемую последовательность, то обнаруженное вхождение для короткого мотива будет достаточно часто иметь высокое (недостоверное) *P*-значение.

На практике длины мотивов сайтов связывания факторов транскрипции высших эукариот находятся в диапазоне 10-20 нуклеотидов (медиана длин – около 12 [Kulakovskiy и др., 2016]). Но для некоторых семейств или регуляторных систем, например, для факторов транскрипции раннего развития *Drosophila*, характерный диапазон длин смещен в сторону коротких мотивов (6-10) и, как следствие, слабозначимых *P*-значений (по сравнению с предсказаниями сайтов других факторов транскрипции). В этом есть и простая логика: в регуляторной системе генов раннего развития ключевую роль играют гомотипические кластеры [Lifanov и др., 2003], и важен именно ансамбль сайтов, а не единичные вхождения.

Анализ значимости множественных вхождений сайтов в последовательности требует своего подхода. Для анализа значимости гомотипических кластеров раньше [Papatsenko, Makeev, 2002] часто использовалась *Z*-оценка (нормализованное на дисперсию отклонение от среднего), но нам кажется более удобным использование обобщенных *P*-значений. Определим обобщенное *P*-значение для протяженной последовательности как вероятность обнаружить не менее чем заданное (наблюдаемое) число вхождений мотива в случайной последовательности фиксированной длины, соответствующей заданной фоновой модели (например, в наиболее простом случае, модели Бернулли с заданными частотами нуклеотидов). Точное вычисление обобщенных *P*-значений с учетом взаимных перекрытий и самоперекрытий мотивов (т.е. возможных пересечений вхождений одного и разных мотивов) является нетривиальным и часто заменяется приближениями и/или эвристиками, например, в предположении о независимости вхождений [Papatsenko, 2007]. Это имеет смысл для высоких порогов оценок и длинных последовательностей, когда вхождения редко перекрываются, а также для конкретных факторов транскрипции, которые склонны связывать единичные сайты.

Точное решение для нескольких сильно перекрывающихся мотивов используется в программе AhoPro [Boeva и др., 2007] на основе суффиксного дерева Ахо-Корасика. Идея получила развитие с использованием графа перекрытий в работе [Régnier и др., 2014] и программе SufPref для сложных фоновых моделей последовательности, но одного мотива. Узкое место точных решений – объем памяти, требуемый для разворачивания мотива как набора допустимых слов в структуру данных для учета перекрытий. На практике оценка обобщенных *P*-значений все еще остается вычислительно трудоемкой, но SufPref, как наиболее продвинутый метод, уже способен адекватно обрабатывать ПВМ длины 15 с реалистичными пороговыми значениями, т.е. анализ точных обобщенных *P*-значений возможен для гомотипических кластеров сайтов связывания большинства факторов транскрипции. Что касается гетеротипических кластеров из сайтов связывания, определяемых различными мотивами, AhoPro успешно справляется с короткими сайтами системы раннего развития *Drosophila* (см., например, ландшафт обобщенных *P*-значений для двух мотивов у [Boeva, 2016]). Точных инструментов, пригодных для длинных и сильно перекрывающихся паттернов пока еще нет.

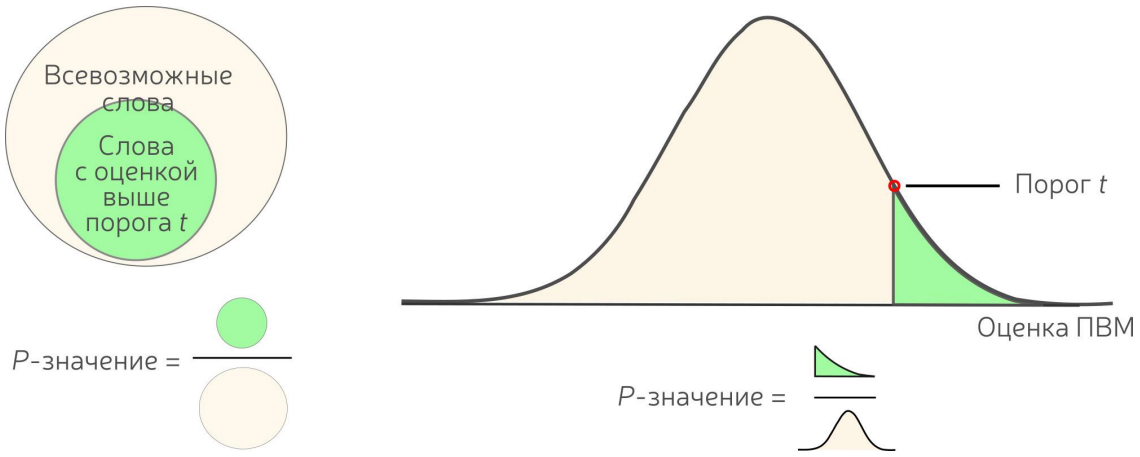


Рисунок 6. *P*-значение мотива для некоторого порога оценки *t*, который выделяет фиксированное подмножество слов с надпороговыми оценками.

3.2.4.2. Мотив как классификатор

Выбор порогового значения оценки превращает модель мотива в бинарный классификатор «да/нет»: принадлежит ли конкретный олигонуклеотид множеству сайтов связывания. В отличие от консенсусных методов, эта классификация не является статической, и ее качество в смысле распознавания истинных сайтов связывания среди всех возможных олигонуклеотидов может быть оценено количественно на различных порогах оценок.

Достаточно часто для сравнения мотивов в смысле точности распознавания сайтов используются стандартные инструменты машинного обучения и анализа сигналов, например кривая операционной характеристики приемника (receiver operating characteristic, ROC). ROC-кривая строится в координатах «доля истинных положительных предсказаний» (ось Y, доля правильно предсказанных сайтов среди успешно верифицированных олигонуклеотидов, чувствительность) и «доля ложных положительных предсказаний» (ось X, доля ошибок первого рода, т.е. ошибочно предсказанных сайтов, среди негативно верифицированных олигонуклеотидов, 1 – специфичность). В качестве свободного параметра выступает пороговое значение оценки, т.е. каждая точка на ROC-кривой соответствует конкретному порогу деления олигонуклеотидов на «сайты» и «не сайты». Пример ROC-кривых для нескольких альтернативных мотивов приведен на **Рисунке 7**. ROC-кривая осмысленного классификатора находится выше диагонали; ROC-кривая хорошего классификатора стремится находится как можно ближе к левому верхнему углу графика (высокая доля истинных и низкая доля ложных положительных предсказаний). Площадь под ROC-кривой (area under ROC-curve, AUC ROC) удобно использовать в качестве количественного значения качества классификации во всем диапазоне пороговых значений.

Универсальность – важное достоинство ROC и аналогичных подходов [Fawcett, 2006; Grau, Grosse, Keilwagen, 2015]. ROC-кривые в случае сайтов связывания пригодны для сравнения различных типов моделей мотивов между собой. Однако, применение ROC в анализе мотивов сталкивается с рядом трудностей и часто критикуется.

Первая проблема: не до конца очевиден объект классификации. Идет ли речь о классификации коротких олигонуклеотидов (точных сайтов) или же геномных

районов связывания (содержащих сайт)? Ряд экспериментальных методов не указывает на точное положение сайта связывания внутри относительно широкого района. Если модель оценивает конкретный олигонуклеотид, а классификации подвергаются протяженные районы, то с ростом длины района растет и эффективное число попыток классификации составляющих их олигонуклеотидов (короткой фиксированной длины – длины мотива) и, следовательно, абсолютное число ложноположительных предсказаний сайтов для каждой тестируемой последовательности. В пределе – в бесконечно длинных последовательностях – штучные истинные сайты будут незаметны на фоне массы случайных ложноположительных предсказаний. То есть, любая достаточно длинная последовательность будет классифицирована положительно.

Современные экспериментальные методы имеют уже достаточное разрешение, чтобы избежать крайностей. Тем не менее, эффективно достижимая площадь под ROC-кривой все равно заметно зависит не только от моделей мотивов, но и от длины последовательностей, которые используются в качестве тестовых выборок (т.е. подвергаются классификации).

Таким образом, для широких районов связывания недостаточно рассматривать единственное наилучшее предсказание единичного сайта, и нужны меры обогащенности, учитывающие множественные вхождения мотива [Kibet, Machanick, 2015], по аналогии с статистикой для гомотипических кластеров: от простой суммы оценок предсказанных сайтов до обобщенного *P*-значения. Однако, на практике одна наилучшая (максимальная) оценка мотива для всех олигонуклеотидов из геномного интервала все еще часто используется [Kulakovskiy и др., 2016; Siebert, Söding, 2016].

Вторая проблема ROC-кривых для мотивов: отсутствие настоящей негативной выборки. В большинстве случаев из эксперимента известны последовательности олигонуклеотидов либо геномные районы, специфически узнаваемые белком. Эти последовательности могут использоваться для подсчета «истинных положительных» предсказаний. Но как подсчитать ложноположительные предсказания? Можно сэмплировать геномные районы поблизости от районов с сайтами связывания [Agius и др., 2010], но трудно удостовериться, что в них действительно отсутствуют функциональные сайты для того же белка. Другой метод – использование набора случайных последовательностей, сгенерированных какой-либо фоновой моделью

[Keilwagen, Grau, 2015; Siebert, Söding, 2016], или вычисление вероятности случайного нахождения сайта связывания для всевозможных последовательностей заданной длины [Kulakovskiy и др., 2013]. По-настоящему эта проблема решается или штучным экспериментальным тестированием кандидатных сайтов [Levitsky и др., 2014] (что помогает и выбору порога распознавания), или высокопроизводительным экспериментальным анализом полного словаря олигонуклеотидов (см. также следующий раздел об экспериментальных методах).

Третья проблема: шкала ROC-кривой показывает весь диапазон ложноположительных предсказаний, тогда как в реальности – с биологической точки зрения – имеют смысл только классификации с крайне низкой долей ложноположительных предсказаний. Для решения этой проблемы предлагается использовать или частичную площадь под ROC-кривой (partial AUC, pAUC [Siebert, Söding, 2016]), или логарифмическую шкалу по оси X (т.е. для доли ложноположительных предсказаний [Medina-Rivera и др., 2011]).

Наконец, в случае, когда истинные «положительные» объекты составляют малую долю от общего множества, что во-многом верно для мотивов, которые описывают лишь малую часть олигонуклеотидного словаря, ряд авторов рекомендует вместо ROC- использовать PR-кривые (PR, precision-recall – точность-полнота) [Davis, Goadrich, 2006; Jankowski, Tiuryn, Prabhakar, 2016].

Вообще, интерпретация сайтов связывания с точки зрения бинарной классификации не очень корректна: фактор транскрипции не имеет жесткого «порога распознавания сайтов» и связывает различные олигонуклеотиды в широком диапазоне аффинностей. К сожалению, систематическая оценка аффинности для различных олигонуклеотидов пока возможна только *in vitro*, и при анализе *in vivo* данных сравнительная оценка качества мотивов остается в рамках задачи о классификации.

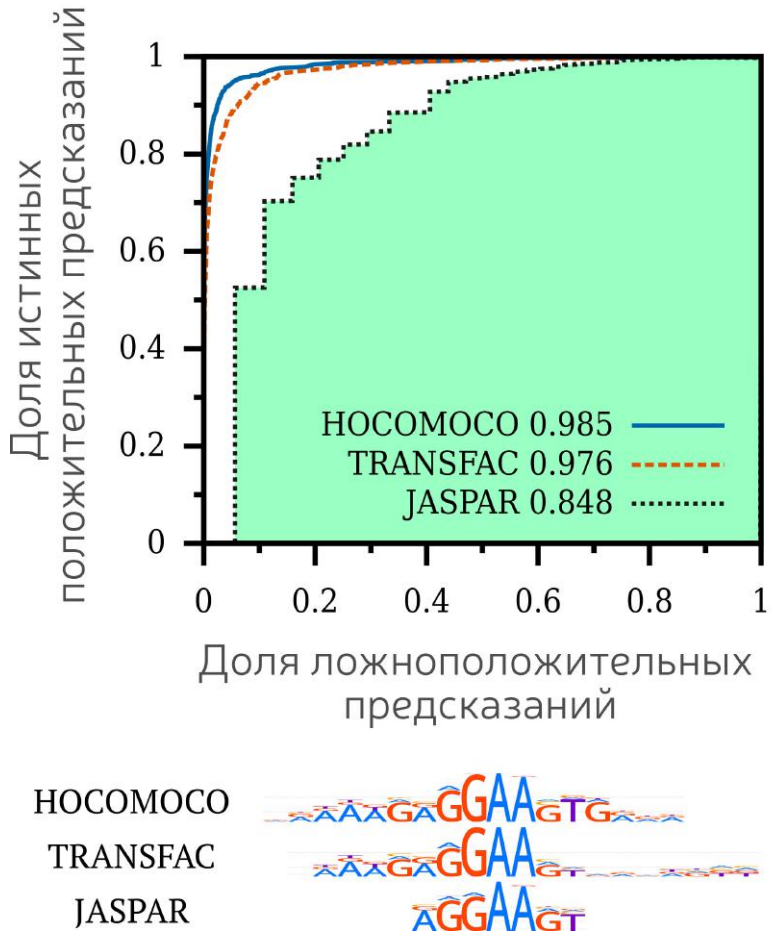


Рисунок 7. ROC-кривые для мотивов связывания фактора Spi1, взятых из публичных (JASPAR, HOCOMOCO) и проприетарных (TRANSFAC) коллекций.

Независимый набор ChIP-Seq данных ENCODE использован в качестве контрольной положительной выборки. Значения площадей под кривыми даны на рисунке. Зеленым цветом выделена площадь под кривой для модели JASPAR. Доля ложноположительных предсказаний оценена по случайным последовательностям. Лого-визуализации мотивов даны на нижней панели.

3.2.4.3. Меры сходства мотивов

Помимо сравнения мотивов с точки зрения точности классификации, не менее интересно, насколько схожи сами паттерны, то есть, описываемые ими множества слов. Это важно и для сравнения результатов различных программ по идентификации мотивов, и для понимания различий в экспериментальных методиках по определению сайтов связывания, и для понимания ожидаемой согласованности предсказаний между различными моделями. Было предложено множество идей и программных реализаций, по-разному решающих задачу сравнения и сопутствующую ей задачу иерархической кластеризации мотивов [Gupta и др., 2007; Habib и др., 2008; Kankainen, Löytynoja, 2007; Mahony, Benos, 2007; Roepcke и др., 2005; Schones, Sumazin, Zhang, 2005; Stegmaier и др., 2013]. Наиболее простые подходы при сравнении весовых матриц напрямую сравнивают колонки выровненных друг относительно друга матриц, например путем подсчета коэффициента корреляции Пирсона [Gupta и др., 2007] или расстояния Кульбака-Лейблера [Roepcke и др., 2005]. Позднее появились и более экзотические подходы, например, сравнение векторов оценок ПВМ для полных наборов олигонуклеотидов фиксированной длины [Xu, Su, 2010].

С биологической точки зрения, логичнее выглядит сравнение ПВМ не как матриц и их элементов, а как моделей мотивов, т.е. сравнение самих предсказываемых сайтов связывания как вхождений мотивов в нуклеотидных последовательностях. Поскольку ПВМ чаще всего используется с фиксированным порогом на распознавание сайта, большие отрицательные значения в матрице, соответствующие неподходящим нуклеотидам, почти никогда не участвуют в предсказаниях сайтов (но, очевидно, учитываются при любом «колоночном» сравнении). Более того, ПВМ, основанные на выравнивании сайтов, не учитывают информацию о последовательностях, имеющих слабую аффинность к изучаемому белку, то есть чаще всего информация о недопредставленных нуклеотидах в ПВМ основана на псевдоотсчетах и является менее достоверной, чем информация о предпочитаемых нуклеотидах. Таким образом, формируется идея «естественных» мер сходства мотивов. Первая подобная мера была предложена в программе MoSta [Pape, Rahmann, Vingron, 2008], разработанной для матриц позиционных отсчетов (или частот до преобразования в веса) на основе ковариации вхождений мотивов в

случайную последовательность ДНК. На практике ПВМ могут быть получены из различных источников с различным весовым преобразованием, различными значениями псевдо-отсчетов и, как следствие, любыми действительными значениями элементов [Levitsky и др., 2007; Stormo, 2000]. Возникает практическая задача по подсчету альтернативной меры сходства, основанной на пересечении множеств слов, подходящих мотиву (например, имеющих оценку ПВМ выше заданного порогового значения). Для сходства мотивов мы предложили использовать меру Жаккара [Vorontsov, Kulakovskiy, Makeev, 2013], которая по постановке не предъявляет конкретных требований к модели. На практике идея была реализована в программе MACRO-APE[17], которая способна сравнивать и ПВМ, и динуклеотидные ПВМ на различных уровнях пороговых значений. Описанию программы посвящен отдельный подраздел «Материалов и методов».

Еще одно направление в сравнении мотивов – поиск ПВМ, похожих на заданную, по коллекции известных мотивов [Bailey и др., 2009; Tanaka и др., 2011]; мера Жаккара может успешно использоваться и в этом случае[18].

3.2.4.4. Аннотация генетических вариантов в некодирующих областях

Помимо предсказания сайтов связывания в регуляторных областях и определения генов-мишеней, мотивы сайтов связывания имеют и другие практические приложения. Научным сообществом уже накоплен огромный объем информации по генетическим вариантам, в первую очередь однонуклеотидным полиморфизмам, локализованным в различных областях генома, и их ассоциации с различными заболеваниями [Eicher и др., 2015]. Заметная доля полиморфизмов локализуется в некодирующих сегментах генома. Нуклеотидные варианты в кодирующей области потенциально могут влиять на структуру и функцию белка, в свою очередь варианты в промоторах и энхансерах могут быть вовлечены в регуляцию транскрипции. Например, многие полиморфизмы, ассоциированные с диабетом второго типа, влияют на связывание фактора RFX [Varshney и др., 2017].

---

[17] MACRO-APE: MAtrix CompaRisOn by Approximate P-value Estimation. http://opera.autosome.ru/macroape/

[18] MACRO-APE: Scan a collection for TFBS models (motifs) similar to a given query. http://opera.autosome.ru/macroape/scan/new

Знание мотивов сайтов связывания позволяет проводить функциональную аннотацию однонуклеотидных вариантов методами биоинформатики. За последние годы для этого была разработана масса инструментов [Macintyre и др., 2010; Manke, Heinig, Vingron, 2010; Riva, 2012; Teng и др., 2012]. В ранних работах предлагалось учитывать сам факт попадания однонуклеотидного варианта в предсказанный сайт [Ponomarenko и др., 2001], но при этом терялась информация об альтернативных аллелях. Более продвинутые инструменты оценивали разницу в предсказанной аффинности сайтов связывания [Manke, Heinig, Vingron, 2010] и статистическую значимость этого изменения [Macintyre и др., 2010]. Наконец, современные методы, помимо оценки изменения аффинности, для выделения наиболее важных полиморфизмов учитывают и существующую геномную аннотацию [Khurana и др., 2013]. На этапе анализа мотивов большинство инструментов использует классические ПВМ. В ходе этой диссертационной работы мы разработали PERFECTOS-APE[19] [Vorontsov и др., 2015], программу для анализа влияния однонуклеотидных замен на аффинность сайтов связывания с использованием как классических ПВМ, так и динуклеотидных весовых матриц. Применение этой программы для анализа соматических мутаций в некодирующих областях генома раковых клеток обсуждается ниже в разделе «Результаты и обсуждение». Характерный пример однонуклеотидного полиморфизма, затрагивающего сайт связывания, проиллюстрирован на **Рисунке 8**.

[19] PERFECTOS-APE: PrEdicting Regulatory Functional Effect by Approximate P-value Estimation. http://opera.autosome.ru/perfectosape/

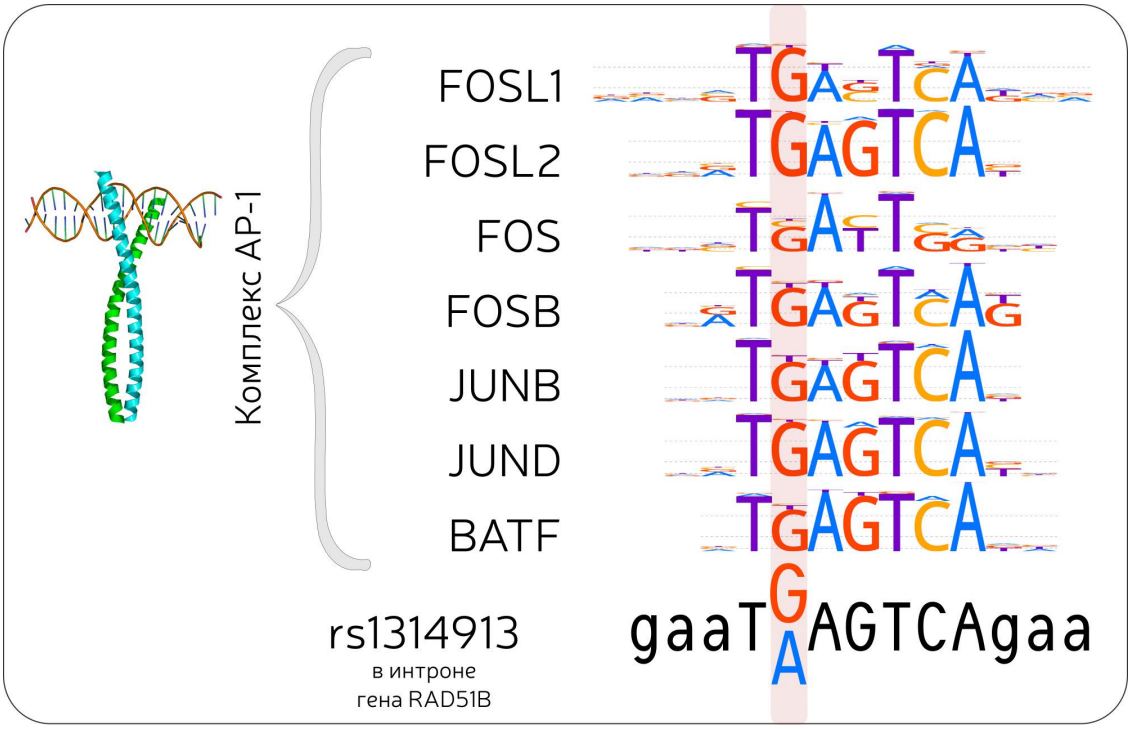


Рисунок 8. Однонуклеотидный полиморфизм rs1314913 локализован в интроне гена RAD51B и влияет на связывание комплекса AP-1.

Белок RAD51 участвует в репарации двухцепочечных разрывов. Аллель T(A) соответствует повышенной в 1.6 раз вероятности развития рака молочной железы у мужчин [Orr и др., 2012]. Аллель G соответствует сильному сайту посадки комплекса-гетеродимера AP-1. Лого-визуализации для мотивов связывания различных членов комплекса приведены по базе HOCOMOCO. Смена аллельного варианта приводит к изменению *P*-значения в 40-280 раз в зависимости от мотива. Рисунок адаптирован из работы [Vorontsov и др., 2015].

3.3. Экспериментальный анализ ДНК-белкового узнавания

Прочтение биологических последовательностей уже перешло из научных проблем в разряд инженерных. Развитие технологий оказалось столь стремительным, что даже в пределах постгеномной эры успел произойти сдвиг парадигмы. Можно провести аналогию между переходом от гибридизационных микрочипов к секвенированию и приходом цифровой эпохи: работа клеточной инженерии теперь видится через дискретные данные, получаемые путем подсчета коротких прочтений фрагментов ДНК и РНК.

В данной работе речь идет о регуляторных мотивах, в первую очередь о паттернах сайтов связывания факторов транскрипции. Биоинформатические методы анализа мотивов развивались параллельно с экспериментальными методами анализа ДНК-белковых взаимодействий еще в «догеномную» эру. Результаты классических экспериментальных методов, безусловно, остаются важными и сегодня, а сами классические методы становятся базой для разработки новых подходов. Мы кратко обсудим и классические экспериментальные методы, и методы на основе гибридизации на микрочипах, и методы на основе высокопроизводительного секвенирования.

3.3.1. Догеномные и постгеномные методы анализа ДНК-белковых взаимодействий

Для анализа ДНК-белкового узнавания на уровне последовательностей ДНК разработан и применяется широкий спектр экспериментальных методов, которые можно условно разделить на две группы: «догеномные» низкопроизводительные классические методы и высокопроизводительнные «постгеномные» методы. Классические низкопроизводительные методы не следует списывать со счетов: несмотря на ограничения «штучного» анализа, они успешно использовались для широкого спектра факторов транскрипции, позволили накопить заметный объем тщательно курированной информации о ДНК-белковом узнавании и продолжают использоваться сегодня, в том числе для верификации высокопроизводительных данных.

Один из первых классических методов – ДНКазный футпринтинг [Galas, Schmitz, 1978], позволяющий локализовать сайт связывания в пределах конкретного изучаемого сегмента ДНК. Меченые по одному концу двуцепочечные фрагменты ДНК, как правило, расщепляют с помощью неспецифической эндонуклеазы (хотя возможны и другие способы внесения неспецифических разрывов в двойную спираль). Полученные фрагменты разделяют с помощью электрофореза и детектируют с помощью радиоавтографии. Фактор транскрипции экранирует от расщепления конкретный участок, соответствующий сайту связывания; в результате на радиоавтографии отсутствуют полосы, соответствующие положению сайта связывания. Разумеется, на основе одного «футпринта» невозможно оценить мотив связывания ТФ; кроме того реагенты для расщепления ДНК (например, типично используемая эндонуклеаза I) в реальности имеют, пусть и слабую, специфичность к локальному нуклеотидному составу, а защищаемая от расщепления область может оказаться шире сайта связывания [Hampshire и др., 2007]. Базовый метод футпринтинга был разработан для использования *in vitro*, но позднее появилась и модификация для анализа *in vivo* [Dai и др., 2000; Mueller, Wold, 1989].

Для построения модели мотива связывания одиночного футпринта недостаточно. Более поздние методы получали информацию о множестве последовательностей в ходе одного эксперимента. Так, техника RSS (random site selection, [Oliphant, Brandl, Struhl, 1989; Blackwell, Weintraub, 1990]) использовала библиотеки меченых случайных олигонуклеотидов, которые инкубировали с целевым белком и разделяли в полиакриламидном геле или с помощью аффинной хроматографии. В результате применения RSS удавалось определить десятки последовательностей, содержащих сайты связывания исследуемого белка. Логичным развитием RSS стал метод SELEX (Systematic Evolution of Ligands by EXponential Enrichment [Tuerk, Gold, 1990]), в котором отбор аптамеров (подходящих по заданному признаку олигонуклеотидов, от лат. aptus – подходить) происходит итеративно в цикле «амплификация – отбор», а для отбора обычно используется аффинная хроматография. За счет циклического отбора SELEX определяет аптамеры с высокой аффинностью к исследуемому белку, в результате обычно удается выявить несколько десятков олигонуклеотидов – сильных сайтов связывания [Goodman и др., 1999; Fields и др., 1997].

Еще один важный классический метод – сдвиг задержки в геле, «гель-шифт» (electro-mobility shift assay, EMSA). Идея состоит в том, чтобы оценить аффинность белка к изучаемому олигонуклеотиду путем оценки изменения мобильности комплекса ДНК-белок в полиакриламидном или агарозном геле. Важная особенность метода – возможность получения количественных оценок аффинности относительно неспецифично связываемого контрольного олигонуклеотида, что позволяет использовать EMSA для детальной верификации вычислительных моделей сайтов связывания и результатов высокопроизводительных методов [Levitsky и др., 2014].

Можно считать, что особенностью футпринтинга является возможность уточнить положение сайта связывания в заданном сегменте ДНК, SELEX определяет набор олигонуклеотидов, которые могут служить в роли наиболее достоверных сайтов связывания, а EMSA дает количественную оценку аффинности-«силы» сайта связывания. При этом, с точки зрения анализа мотивов, итоговый набор последовательностей, получаемых «догеномными» методами, имеет небольшой масштаб: от считанных штук до малых десятков последовательностей длиной в десятки нуклеотидных оснований. Информация о сайтах связывания, определенных классическими методами, систематизирована только ограничено. За вычетом ресурсов, сфокусированных на конкретном виде или типе данных (например, footprintDB для футпринтинга сайтов связывания факторов транскрипции *D. melanogaster* [Bergman, Carlson, Celniker, 2005]), полноценно охватывала факторы транскрипции высших эукариот только коммерческая база данных TRANSFAC [Matys и др., 2003]. Это заметно ограничивало возможности систематических исследований *in silico*, но стало менее критично с появлением данных массовых высокопроизводительных методов.

Современные высокопроизводительные методы можно базово разделить на три группы: *in vitro* с искусственными олигонуклеотидами, *in vitro* с геномными фрагментами ДНК, и методы *in vivo*. Начнем рассказ с наиболее успешного метода *in vitro* под названием PBM (protein-binding microarray [Mukherjee и др., 2004; Berger и др., 2006]), использующем гибридизацию на микрочипах (как и похожий метод CSI, cognate site identifier [Warren и др., 2006]). PBM, пожалуй, единственный «гибридизационный» метод по анализу сайтов связывания, который не уступил позиций с приходом массового секвенирования. Для анализа сайтов связывания PBM

использует микрочипы, на которых закреплены различные синтетические двуцепочечные олигонуклеотиды. Каноничный метод дизайна т.н. универсальных белок-связывающих микрочипов (universal PBM [Philippakis и др., 2008]) предполагает расположение на подложке олигонуклеотидов из 60 пар оснований, которые в сумме содержат вхождения всевозможных 10-нуклеотидных подстрок. Для оценки связывания исследуемый белок флюоресцентно метится (либо сшивается с эпитопом, для которого существуют меченые антитела) и инкубируется с микрочипом. С анализом гибридизационных данных есть ряд типичных сложностей, связанных с насыщением флюоресцентного сигнала и неспецифической гибридизациацией. Тем не менее, подход PBM достаточно уникален, поскольку предоставляет количественную информацию о связывании и не-связывании белком широчайшего спектра олигонуклеотидов (в пределе покрывая полный «словарь» ДНК-подстрок фиксированной длины). Здесь критически важно, что в результате присутствует и систематическая информация о слабых сайтах и «не сайтах», т.е. экспериментально-обоснованный негативный контроль. Мотивы связывания сотен факторов транскрипции для различных организмов, определенные с помощью PBM, представлены в базе данных UniProbe [Newburger, Bulyk, 2009; Hume и др., 2015]. В последние годы область применения PBM расширилась: появляются исследования композитных элементов сайтов связывания для посадки белковых комплексов [He и др., 2015; Siggers, Gordân, 2014].

Еще один интересный метод *in vitro* , использующий микрочипы – DIP-chip (DNA-immunoprecipitation chip [Liu и др., 2005]). В этом случае пробы на микрочипе соответствуют реальным участкам изучаемого генома. Очищенная ДНК из целевого организма фрагментируется ультразвуком, полученные фрагменты инкубируют с исследуемым белком. С помощью иммунопреципитации за специфические антитела выделяется связавшая белок фракция ДНК, которая метится и конкурентно с геномной ДНК гибридизуется на микрочипе. Фактически метод работает *in vitro*, но определяет сайты не в случайных олигонуклеотидах, а в реальных геномных последовательностях. Эта же идея – использование реальной геномной ДНК *in vitro* для поиска геномных сайтов связывания – используется и в более современных методах на основе высокопроизводительного секвенирования, например DIP-Seq [Gossett, Lieb, 2008], PB-Seq [Guertin и др., 2012] и сравнительно молодом методе

DAP-Seq [O'Malley и др., 2016]. Важная особенность, которую можно рассматривать и как достоинство и как недостаток, – анализ *in vitro* не учитывает структуру укладки хроматина и доступность конкретных сайтов для связывания в конкретном типе клеток. Помимо ухода от микрочипов, рассматриваемые методы имеют еще одну особенность, упрощающую проведение экспериментов для широкого спектра факторов транскрипции, а именно, отсутствует необходимость в антителах на исследуемый белок. Вместо иммунопреципитации за специфические антитела используются рекомбинантные химерные белки, в которых исследуемый фактор транскрипции объединен с известным тагом для последующей очистки связанной фракции ДНК.

Трудности в получении высокоспецифических антител вынуждают к использованию *in vitro* методов или поиску обходных путей. Например, DamID: предлагает определение сайта связывания фактора транскрипции по метилированию ДНК, осуществляемому *in vivo* фьюжном фактора транскрипции и метилтрансферазы Dam [Steensel van, Delrow, Henikoff, 2001], однако грубое разрешение получаемой карты индуцированного локального метилирования не позволяет точно идентифицировать конкретные сайты связывания.

Интересно предсказуемое эволюционное развитие методов с переходом от «штучного» анализа к технологиям глубокого секвенирования: EMSA-Seq [Wong и др., 2011], SELEX-seq [Slattery и др., 2011] (также называемый HT-SELEX [Jolma и др., 2013]), DamID-seq [Wu, Olson, Yao, 2016]. Особняком стоит метод «полногеномного масштабирования» ДНКазного футпринтинга, DNase-Seq [Boyle и др., 2008; Song, Crawford, 2010], предоставляющий информацию о геномных футпринтах всевозможных белков внутри районов доступного хроматина (но не указывающий конкретных белков, которые связывали тот или иной сайт). Сестринский метод – ATAC-seq [Buenrostro и др., 2013] – картирует районы открытого хроматина с помощью транспозиции адаптеров для секвенирования непосредственно в нативную ДНК, и, по сравнению с DNase-seq, требует меньшего объема биологического материала (вплоть до нескольких сотен клеток).

Особое место занимает группа современных методов, предназначенных для количественной оценки аффинности белка к ДНК. Например, в работе [Nutiu и др., 2011] была предложена остроумная методика HiTS-FLIP (high-throughput sequencing

– fluorescent ligand interaction profiling). До некоторой степени она является заменой PBM с секвенированием вместо гибридизации: библиотека случайных олигонуклеотидов прочитывается в проточной ячейке современного высокопроизводительного секвенатора, а затем непосредственно в той же ячейке происходит инкубирование флюоресцентно-меченого белка с отсеквенированными олигонуклеотидами и оптика секвенатора используется для считывания картины флюоресценции белка, т.е. для одновременной оценки связывания белка с миллионами случайных олигонуклеотидов (их последовательности к этому моменту уже прочитаны и известны). Используя различные концентрации белка, удается не только детально изучить мотив связывания, но и оценить аффинность для широчайшей библиотеки олигонуклеотидов, по размеру сравнимой, или даже превышающей таковую для PBM. Также активно развиваются методы на основе микрофлюидики: в комбинации с микрочипами для параллелизации измерения аффинности связывания белка в различных концентрациях с различными олигонуклеотидами (BunDLE-Seq [Levo и др., 2015], MITOMI [Rockel, Geertz, Maerkl, 2012], QPID [Glick и др., 2015]) и в комбинации с последующим глубоким секвенированием [Isakova и др., 2017].

Наконец, нельзя не упомянуть наиболее популярные и востребованные на сегодня высокопроизводительные методы, основанные на иммунопреципитации хроматина *in vivo*: ChIP-chip и ChIP-Seq; они подробно обсуждаются в следующем разделе.

Сравнительный обзор современных высокопроизводительных методов, включая более общие подходы для идентификации открытых сегментов хроматина, свободного от нуклеосом, дан в работе [Levo, Segal, 2014].

Для всех современных экспериментальных методов анализа ДНК-белковых взаимодействий характерны большие объемы данных (тысячи, сотни тысяч и более детектируемых последовательностей). С другой стороны методы отличаются систематическими ошибками, имеют различную разрешающую способность (от считанных нуклеотидов для HT-SELEX до тысяч нуклеотидов в случае DamID) и степень определенности по поводу фактора транскрипции (конкретный белок в случае ChIP-Seq или любой релевантный фактор транскрипции в случае DNase-seq). Анализ мотивов (и идентификация *de novo* и поиск вхождений известных паттернов)

является одним из ключевых шагов в точной идентификации сайтов связывания. Применение мотивов в анализе ChIP-Seq данных обсуждается в следующем разделе.

3.3.2. Анализ полногеномного профиля связывания ДНК факторами транскрипции путем иммунопреципитации хроматина с последующим глубоким секвенированием

Иммунопреципитация хроматина с последующим глубоким секвенированием (ChIP-Seq, Chromatin ImmunoPrecipitation followed by massively parallel/deep Sequencing) – текущий золотой стандарт и наиболее популярный метод исследования связывания ДНК белком *in vivo* [Liu, Pott, Huss, 2010]; это метод, пригодный для исследования геномного ландшафта расположения разнообразных белков, от локализации специфических модификаций гистонов до сайтов связывания факторов транскрипции [Park, 2009]. Для факторов транскрипции ChIP-Seq детектирует тысячи и десятки тысяч участков связывания в геномах высших эукариот. Не совсем корректно говорить, что ChIP-Seq определяет непосредственно сайты связывания белков, поскольку длина детектируемых участков может быть в сотни раз больше единичного сайта. Строго говоря, задача обработки данных ChIP-Seq для определения положения сайта связывания с точностью до отдельных нуклеотидов решена не полностью, но предложено множество подходов.

В этом разделе мы обсудим базовый анализ данных ChIP-Seq и важные особенности эксперимента, которые следует принимать во внимание. Далее мы покажем, как методы анализа мотивов могут использоваться для уточнения локализации сайтов связывания, верификации экспериментальных данных и получения интересных биологических наблюдений о схемах расположения сайтов связывания в регуляторных районах.

Ранние попытки систематизации сайтов связывания и соответствующих мотивов [Heinemeyer и др., 1998; Sandelin и др., 2004; Wingender и др., 2000] основывались на данных «догеномных» низкопроизводительных методов и достаточно часто возможности описания мотива как сходства между сайтами были прямо ограничены считанным числом экспериментально определенных последовательностей. Появление массовых данных ChIP-Seq пробудило интерес к выявлению и более точному представлению мотивов сайтов связывания и, одновременно, именно анализ мотивов в данных ChIP-Seq может стать финальным

аргументом в споре о специфичности связывания конкретного белка *in vivo*. Один из таких примеров – результат анализа мотивов в ChIP-Seq данных многофункционального белка (в том числе, и фактора транскрипции) YB-1 [Astanehe и др., 2012; Dolfini, Mantovani, 2012], показавший, что систематически YB-1 не связывает CCAAT-сайты в ДНК *in vivo*, вопреки устоявшимся мнениям [Ise и др., 1999].

3.3.2.1. От гибридизации к секвенированию: -chip *versus* -Seq

Чтобы выделить характерные особенности ChIP-Seq, которые могут стать как источником статистического шума, так и полезным априорным знанием в анализе ДНК-белкового узнавания, нам пригодится понимание пошагового процесса экспериментального получения (wet lab) и компьютерной обработки данных (dry lab).

Экспериментальная часть метода ChIP-Seq во-многом повторяет предыдущие геномные технологии, использующие гибридизацию на микрочипах [Horak, Snyder, 2002]. Наиболее популярным среди похожих методом стал ChIP-chip (ChIP-on-chip, Chromatin ImmunoPrecipitation + microchip). При всех систематических искажениях, проистекающих от техники гибридизации, ChIP-chip с выстилающими геном микрочипами (tiling arrays, т.е. содержащими пробы, регулярно покрывающие геномную последовательность [Lemetre, Zhang, 2013]) способны показать полногеномную картину связывания *in vivo*; это положительно сказалось и на последующей популярности ChIP-Seq.

Экспериментальную стадию протокола ChIP-chip можно упрощенно разделить на несколько шагов. Сначала производится ковалентная сшивка белков и клеточной ДНК (например, с помощью обработки формальдегидом); затем клетки лизируются и выполняется фрагментация ДНК (обычно используется обработка ультразвуком, «соникация»); сшитые комплексы ДНК-белок экстрагируют и извлекают с помощью специфичных антител (иммунопреципитация); сшивки снимают, выполняют очистку, амплификацию ДНК, флюоресцентное мечение и гибридизацию амплифицированных фрагментов на ДНК-микрочипе. Важная и не совсем очевидная проблема технологии ChIP-chip – необходимость интеллектуального дизайна выстилающих микрочипов. В экспериментах по определению экспрессии генов пробы на микрочипе могут покрывать только малую долю геномных

последовательностей (транскриптом), но для анализа сайтов связывания пробами необходимо покрыть значительно больший процент геномной последовательности (в идеале – весь геном; но для протяженных геномов на практике покрывалась только их часть, например, только промоторы аннотированных генов или только некоторые хромосомы [Cawley и др., 2004]). Количество проб определяет их максимальную плотность в зависимости от размера генома и, следовательно, ограничивает эффективное разрешение при определении сайтов связывания. Кроме того, повторимся, ChIP-chip страдает от проблем, общих для экспериментальных техник на основе гибридизации [Buck, Lieb, 2004], в том числе, в связи с насыщением и шумностью флюоресцентного сигнала. И разумеется, интерпретация результатов процедуры ChIP-chip сильно зависима от вычислительной обработки, в том числе, от анализа мотивов [Ettwiller и др., 2007; Linhart, Halperin, Shamir, 2008].

ChIP-Seq [Robertson и др., 2007] в заметной степени повторяет ChIP-chip в части экспериментального протокола, в котором гибридизация заменяется на высокопроизводительное секвенирование. ChIP-Seq порождает миллионы одно- или двухконцевых прочтений иммунопреципитированных и амплифицированных фрагментов ДНК. Удивительно, что минорное изменение экспериментальной процедуры произвело революционный эффект. ChIP-Seq и похожие «сестринские» технологии [Ng и др., 2006; Wei и др., 2006] стали основной и классических работ по изучению отдельных белков [Johnson и др., 2007; Robertson и др., 2007; Valouev и др., 2008], и крупномасштабных проектов, таких как ENCODE [Dunham и др., 2012], и уже позволяют проводить анализ в нетипичных случаях, например, изучать сайты связывания белков в геномах архей [Wilbanks и др., 2012].

3.3.2.2. ChIP-Seq эксперимент и точность определения сайтов связывания

Грамотно поставленный эксперимент ChIP-Seq может включать этап отбора фрагментов ДНК по размеру (size selection) и контрольные данные, полученные либо путем прочтения фрагментированной геномной ДНК без иммунопреципитации (input DNA, [Johnson и др., 2007]), либо иммунопреципитацией за неспецифичные антитела [Chen и др., 2008], например, с использованием преимунной сыворотки. В обоих случаях, экспериментальная процедура завершается высокопроизводительным секвенированием, порождая одно- или двухконцевые прочтения фрагментов ДНК. Прочтения (чтения, reads, tags) картируются на существующую сборку генома

(reference genome assembly); задача картирования – отдельная сложная тема, см. например обзор [Leleu, Lefebvre, Rougemont, 2010]. В случае ChIP-Seq задача картирования является важной, но не критической, поскольку на следующем шаге картирование используется в процедуре идентификации кластеризованных групп чтений, «пиков» (peak calling), которая достаточно устойчива к локальным флуктуациям (за исключением серьезных артефактов ПЦР и неуникально картируемых чтений, см. также [Chung и др., 2011; Wang и др., 2013b]).

Важно понимать, что чтения только косвенно соответствуют сайтам связывания, которые локализованы где-то в окрестности соответствующих геномных картирований, а размер окрестности зависит от распределения длин фрагментов ДНК, полученных в результате иммунопреципитации и использованных для секвенирования. Кроме того, сами иммунопреципитированные фрагменты скорее всего содержат сайт связывания внутри заметно более длинных фланкирующих последовательностей. Это легко понять из соображения о характерной длине фрагментов ДНК после соникации [Qi и др., 2006] в районе 200 нуклеотидов, что на порядок превышает характерную длину сайтов связывания. Таким образом, в конкретном регионе генома, базовый профиль покрытия чтениями (в английской литературе используются различные термины, в т.ч. reads pileup profile, tag density profile, base coverage profile) только примерно показывает положение фактических сайтов связывания. Процедура идентификации пиков нужна именно для выделения сегментов ДНК, обогащенных картированными чтениями [Rye, Sætrom, Drabløs, 2011; Wilbanks, Facciotti, 2010].

Для выделения пиков возможны две базовых стратегии работы с картированными чтениями. Первый, безопасный путь – «расширение чтений», использовался, например, некогда популярной программой FindPeaks [Fejes и др., 2008]. Идея состоит в том, чтобы расширить закартированные чтения до предполагаемой исходной длины фрагментов ДНК и использовать для поиска сайтов перестроенный профиль на основе покрытия генома «расширенными» фрагментами. В реальности парные чтения для ChIP-Seq используются редко (в первую очередь по финансовым соображениям), а для одноконцевых чтений необходима эвристика, описывающая ожидаемое распределение длин фрагментов. Кроме того, получаемый при расширении прочтений профиль естественным образом сглаживается и итоговые

пики по построению представляют собой «широкие районы связывания белка», а не конкретные сайты связывания. Альтернативная стратегия состоит в раздельном «направленном» рассмотрении чтений, картированных на прямую и обратную цепь конкретного сегмента ДНК, например как в работе [Jothi и др., 2008]. Здесь идея состоит в том, что наличие сайта связывания – это то общее, что есть у иммунопреципитированных фрагментов, чтения которых локализованы в одной области генома. Рассмотрим идеализированную картину: секвенируются 5’ концы фрагментов ДНК; т.е. относительно чтений, картированных на конкретную цепь ДНК, сайт связывания должен быть локализован в 3’ направлении, и 3’ граница сайта должна соответствовать окончанию обогащения фрагментами. Поскольку это верно для обоих цепей ДНК, то можно ожидать «переходной точки», сдвига или региона относительно обедненного чтениями, который и будет соответствовать точной локализации сайта связывания (см. пример картирования на **Рисунке 9**). Логично, что методы, учитывающие направление картирования чтений обладают значительно лучшей разрешающей способностью [Wilbanks, Facciotti, 2010].

Практическая эффективность обоих стратегий идентификации пиков ограничена особенностями реальных регуляторных сегментов генома. Вспомним, что геномные сайты связывания расположены неслучайно и образуют как неструктурированные гомо- и гетеротипические кластеры [Gotea и др., 2010; Lifanov и др., 2003; Murakami, Kojima, Sakaki, 2004], так и структурированные композитные элементы [Matys и др., 2006; Shelest, Albrecht, Shelest, 2010]. Поскольку типичное расстояние между сайтами связывания меньше длины иммунопреципитируемого фрагмента ДНК, картирование чтений является результатом интерференции сигналов от близлежащих сайтов связывания. В случае выраженных единичных сайтов связывания (например, ChIP-Seq для факторов транскрипции прокариот), удается провести декомпозицию пиков и четко различить сигналы от соседних сайтов [Chung и др., 2013; Lun и др., 2009]. Однако для эукариот типичны «тесные» гомотипические кластеры, в которых сайты связывания расположены близко или даже перекрываются [Gotea и др., 2010; Lifanov и др., 2003], что заметно зашумляет картирование. Связанная проблема состоит в том, что различные подмножества сайтов в одном участке ДНК могут быть по-разному заняты белком в различных клетках исследуемой культуры. Это также вносит шум, который может приводить к

неверной аннотации сайтов при попытке уточнить их локализацию с помощью направленного подсчета чтений и ошибкам в автоматической оценке исходной длины фрагментов ДНК (которую проводит, например, популярная программа MACS [Zhang и др., 2008]).

Локализация сайтов связывания может быть установлена недостаточно точно при избыточном расширении пиков и объединении сигналов от нескольких сайтов связывания, либо, для узких пиков, реальный сайт связывания может оказаться в стороне от найденного пика – региона связывания. Для некоторых задач (например, установления списка потенциальных генов-мишеней с сайтами связывания в промоторах) такая точность достаточна. Для других приложений точность может быть улучшена путем использования дополнительных инструментов. В частности, анализ мотивов позволяет идентифицировать сайты связывания с разрешением вплоть до отдельных нуклеотидов.

На **Рисунке 10** представлена пошаговая схема эксперимента и обработки данных ChIP-Seq. Последний этап – использование методов анализа последовательностей для точной идентификации сайтов связывания – находится в фокусе нашей работы.

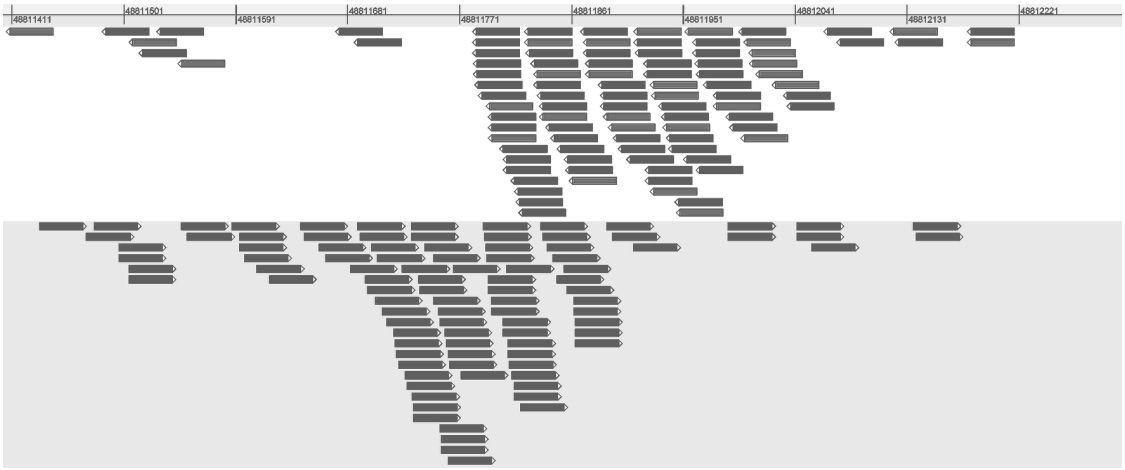


Рисунок 9. Пример характерной асимметричной кластеризации картированных чтений по прямой (*верхняя панель*) и обратной (*нижняя панель*) цепям ДНК. Показан участок 20й хромосомы генома человека, данные ChIP-Seq для белка FoxA1 [Zhang и др., 2008], визуализация с помощью GenomeView [Abeel и др., 2012].

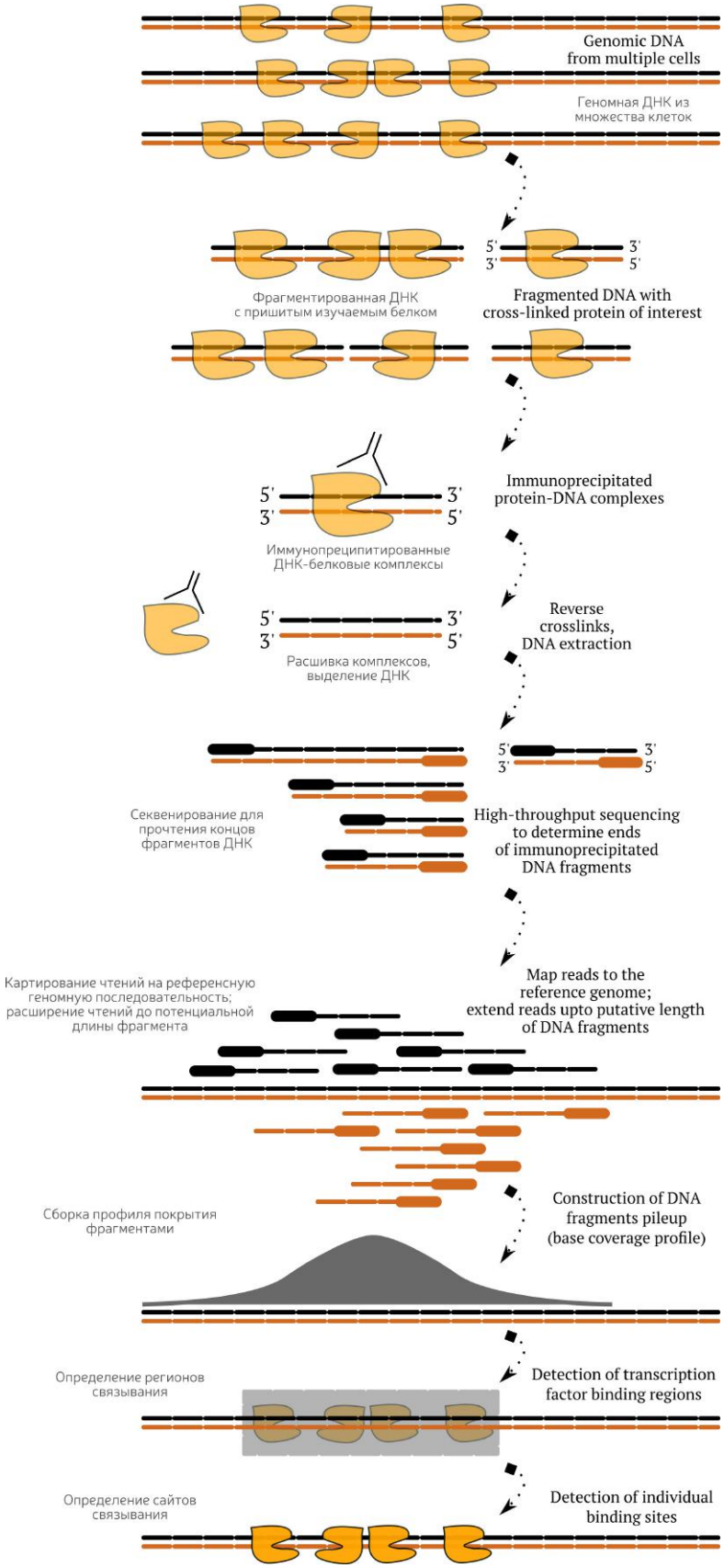


Рисунок 10. Концептуальная схема ChIP-Seq.

Рисунок адаптирован из обзора [Kulakovskiy, Makeev, 2013].

3.3.2.3. Локализация сайтов связывания в пиках

ChIP-Seq предоставляет прямую информацию об участках ДНК, связываемых белками, но без дополнительной вычислительной обработки не позволяет выявить отдельные сайты связывания в протяженных сегментах ДНК и кластерах сайтов. Тем не менее, можно оценить разумность экспериментальных данных даже используя упрощенное предположении о единственном сайте связывания в каждом пике. А именно, сегменты ДНК с лучшими сайтами связывания (т.е. обладающими большей аффинностью к изучаемому белку) должны иметь лучшее покрытие чтениями. То есть лучшие сайты связывания должны быть локализованы в более высоких пиках и обратно. Это проиллюстрировано на **Рисунке 11**, где с помощью существующей ПВМ мы разметили сайты связывания в ChIP-Seq пиках для белков AP2A и E2F1 [Kulakovskiy и др., 2013]. Пики отсортированы по высоте по убыванию, т.е. высокие пики имеют наименьшие ранги. Для AP2A явно виден выраженный тренд (который хуже прослеживается для E2F1): наилучшие оценки ПВМ (указывающие на наилучшие сайты связывания в каждом пике) падают для пиков с большими рангами (т.е. меньшими высотами). В то же время, ранжирование пиков по статистической значимости или высоте коррелирует и с числом сайтов в пиках [Régnier и др., 2014]. К сожалению, картину несколько искажает то, что высокие пики часто длиннее, чем низкие (в силу того, что в высоких пиках могут перекрываться сигналы от нескольких соседних сайтов связывания); а в длинных последовательностях чаще, чем в коротких, будут случайно встречаться и сильные единичные и средние множественные предсказания ПВМ.

В массе опубликованных данных представлены только высоты пиков ChIP-Seq (максимальное число перекрывающихся чтений) или значения статистической значимости для обогащения прочтениями конкретных сегментов ДНК по сравнению с контролем. В принципе, информация о высоте пика и\или обогащении достаточна для определения поднабора наиболее достоверно определенных участков генома, связываемых изучаемым белком. Полногеномные профили покрытия прочтениями не сохраняют из практических соображений – для экономии объема файлов. В то же время, форма пиков (т.е. форма профиля покрытия фрагментами) содержит важную позиционную информацию помимо, собственно, высоты. Важнейшее свойство профилей – локализация вершины пика (peak summit, явным образом сохраняется,

например, при использовании программы MACS [Zhang и др., 2008]). Для ассиметричного пика с неравновесными плечами информация о положении вершины, как участка, наиболее плотно покрытого чтениями, может указывать на вероятное положение наиболее сильного сайта связывания внутри широкого геномного интервала. И обратно, корректная локализация сайтов связывания у вершины пика косвенно свидетельствует в пользу корректности экспериментальных данных.

Идею о локализации сайтов связывания относительно вершин пиков можно проиллюстрировать на примере. Определим относительное положение сайта связывания в пике как d_s/d_b, где d_s это расстояние от сайта (наилучшего предсказания ПВМ) до вершины пика, а d_b это расстояние до ближайшей границы пика (левой или правой). На **Рисунке 12** показано кумулятивное распределение (с накоплением) доли пиков для фактора AP2A в зависимости от d_s/d_b. График демонстрирует, что предсказанные сайты связывания локализуются в окрестности вершин пиков и избегают фланкирующих участков. Важно, что кривая для ошибочной ПВМ (для белка E2F1) ведет себя так же, как кривая для корректной ПВМ (AP2A, на который сделан ChIP-Seq) на последовательностях пиков с перемешанными нуклеотидами. Наблюдение о предпочтительной локализации предсказанных сайтов в районе вершины согласуется с процедурой ChIP-Seq и часто используется на практике, когда для простоты все пики обрезаются на фиксированном расстоянии вокруг вершины. В следующем разделе мы обсуждаем примеры, когда прямолинейная стратегия обрезки пиков бывает ошибочна.

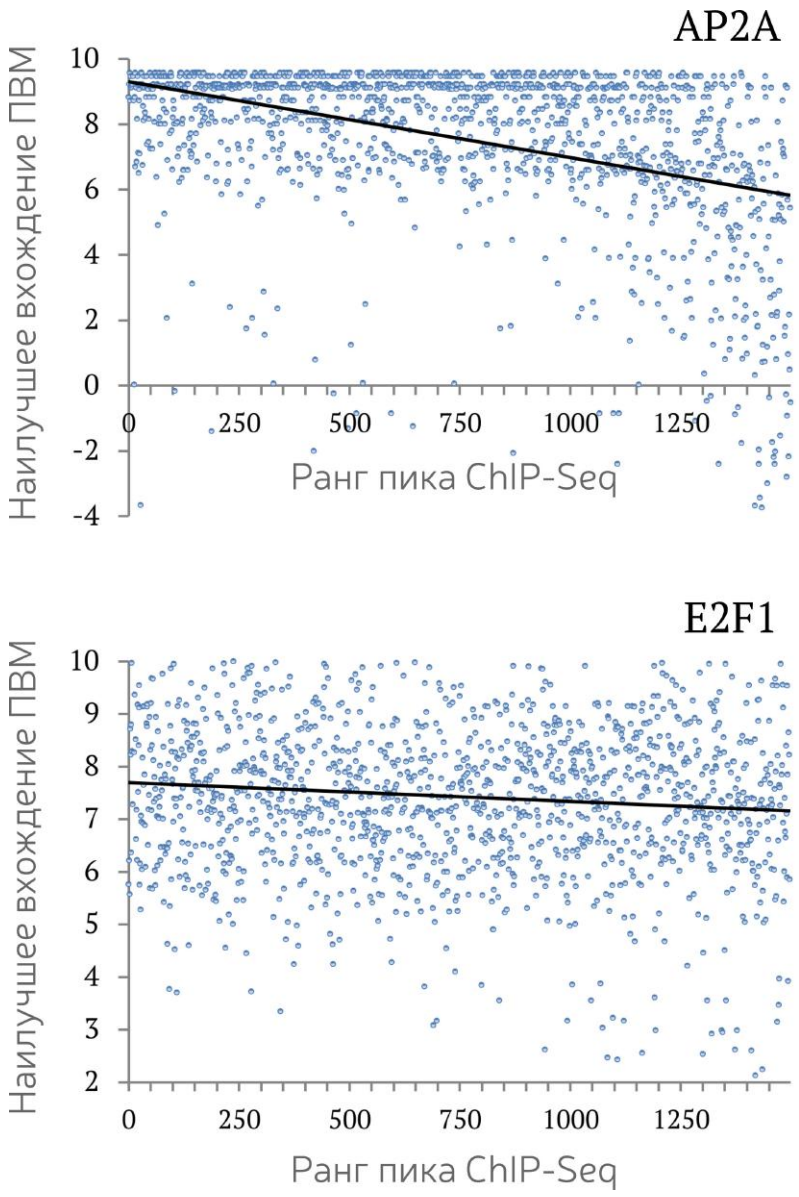


Рисунок 11. Наилучшие вхождения ПВМ для AP2A (*верхняя панель*) и E2F1 (*нижняя панель*) в ранжированные по высоте пики ChIP-Seq (по данным ENCODE).

Ось X соответствует рангам пиков, ось Y – наилучшей оценке ПВМ для конкретного пика. Каждый пик показан отдельной точкой, сплошная линия показывает линейный тренд. Рисунок адаптирован из обзора [Kulakovskiy, Makeev, 2013].

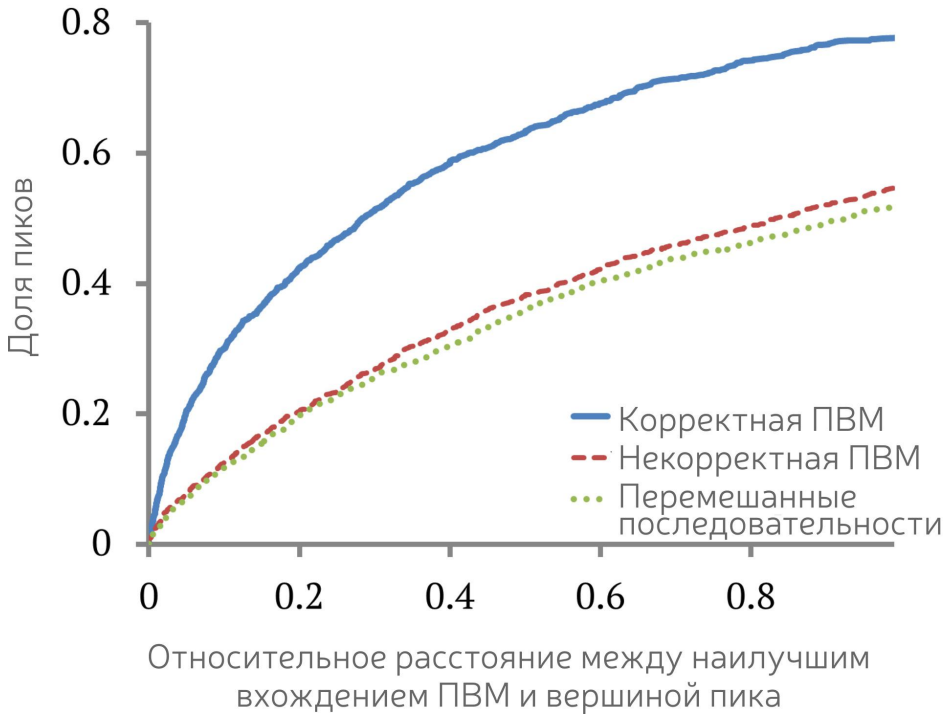


Рисунок 12. Кумулятивное распределение доли пиков AP2A (ось Y) для которых относительное расстояние между наилучшим вхождением ПВМ и вершинами не превосходит порога (ось X).

Корректная ПВМ соответствует сайтам AP2A, некорректная – сайтам E2F1. Рисунок адаптирован из обзора [Kulakovskiy, Makeev, 2013].

3.3.2.4. Особенности формы пиков

Положение вершин в пиках коррелирует с локализацией сайтов связывания. Рассмотрим, как выглядит форма реальных пиков ChIP-Seq при реконструкции покрытия фрагментами ДНК. Если в различных клетках одного образца фрагментация ДНК случайна, а глубина секвенирования достаточна, можно ожидать гладкой колоколообразной формы пика вокруг сильного единичного сайта связывания либо группы близких сайтов. На практике ситуация несколько сложнее. **Рисунок 13** демонстрирует трио «хороший, плохой, злой» из высоких пиков для (1) классического эксперимента с белком GABPA [Valouev и др., 2008], (2) эксперимента с онкогенным фьюжн-белком EWS-FLI1 [Guillon и др., 2009] и (3) эксперимента для белка AP2A из ранних данных проекта ENCODE [Myers и др., 2011]. Пик для GABPA имеет характерную форму колокола, профиль пика для EWS-FLI1 изрезан и для него «обрезка вокруг вершины» выглядит неприменимой. Данные для AP2A, по всей видимости, демонстрируют артефакт картирования или ПЦР-амплификации (множество идентичных прочтений, картированных на один и тот же участок генома). Эта проблема частично решается дедупликацией чтений, например, с помощью вычисления вероятности случайного сэмплирования идентичных чтений при заданной глубине секвенирования, которую некоторые методы идентификации пиков выполняют автоматически. Однако на практике «квадратные» пики все равно иногда попадают в список наиболее высоких/значимых и, без дополнительной фильтрации и/или ручного контроля, вносят шум в последующий анализ.

3.3.2.5. Эффект гомотипической кластеризации сайтов связывания в пиках

Высокие пики классической колоколообразной формы значительно длиннее, чем иммунопреципитированные фрагменты ДНК, а ширина «колокола» часто заметно больше размера одиночного сайта связывания. Это может объясняться сложением сигналов от соседних сайтов связывания в одном регуляторном районе, где изучаемый белок в различных клетках связывает близкие, но не одни и те же сайты в силу стохастических причин. Проиллюстрировать это соображение можно с помощью поиска вхождений мотивов. В качестве тематического исследования подходят регуляторные районы генов, участвующих в раннем развитии плодовой мушки *Drosophila melanogaster*, для которых хорошо выражен феномен гомотипической кластеризации. Рассмотрим локус гена *giant* (**Рисунок 14**). Здесь

используются ChIP-Seq данные для известных регуляторов Bicoid, Caudal и Hunchback на 5 стадии раннего развития [Kaplan и др., 2011], наложена информация о ДНКазной доступности [Li и др., 2011a] и предсказания сайтов связывания с помощью ПВМ из работы [Kulakovskiy, Makeev, 2009], а в качестве изучаемого локуса взят регион (-12000; +6000) относительно аннотированного сайта инициации транскрипции.

Как мы уже отмечали, факторы транскрипции, участвующие в регуляции транскрипции в раннем развитии *Drosophila*, связывают плотные гомотипические кластеры сайтов связывания (Lifanov et al., 2003). Для построения трека значимости гомотипических кластеров, представленного на **Рисунке 14**, мы использовали программу AhoPro [Boeva и др., 2007] для подсчета *P*-значений как вероятностей обнаружения кластера сайтов связывания в случайной последовательности. Говоря формально, в качестве статистической значимости кластеров подсчитана вероятность p наблюдения не менее чем заданного числа вхождений мотива в скользящем окне (50 пар оснований) с учетом локального нуклеотидного состава (т.е. относительно фоновой модели Бернулли с оцененными в окне частотами нуклеотидов). Профиль статистической значимости гомотипической кластеризации $-\log(p)$ показан в качестве отдельной дорожки. В качестве иллюстративного примера мы выбрали ген *giant*, но схожие эффекты наблюдаются и для других генов раннего развития, например *hunchback* и *knirps*. Удивительно, что если исключить ложноположительные предсказания ПВМ вне ДНКазо-доступных районов, форма пиков ChIP-Seq хорошо согласуется с профилем гомотипической кластеризации. При этом, мы никак не оптимизировали параметры, такие как размер окна и фиксированный порог на распознавание ПВМ [Kulakovskiy, Makeev, 2009]. То есть, схема расположения сайтов связывания в заметной степени определяет форму пиков ChIP-Seq. И, как уже говорилось, для выраженных отдельных сайтов связывания (например, у прокариот) это позволяет успешно проводить декомпозицию для распознавания сигнала от соседних сайтов связывания [Lun и др., 2009].

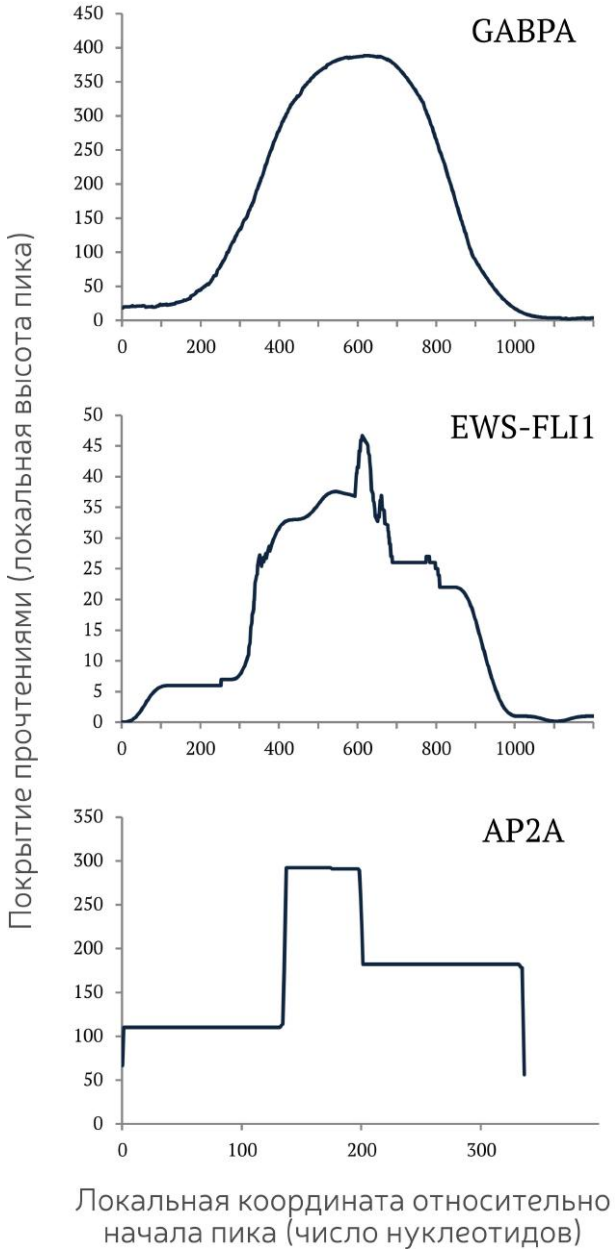


Рисунок 13. Примеры формы пиков ChIP-Seq для наиболее высоких пиков из трех независимых экспериментов для разных факторов транскрипции.

Показаны пики для GABPA (*верхняя панель*, классическая колоколообразная форма), EWS-FLI1 (*средняя панель*, относительно шумный профиль), AP2A (*нижняя панель*, потенциальный артефакт, вызванный дуплицированными прочтениями). Ось Y показывает глубину покрытия в относительной позиции сегмента пика, отложенной по оси X. Рисунок адаптирован из обзора [Kulakovskiy, Makeev, 2013].

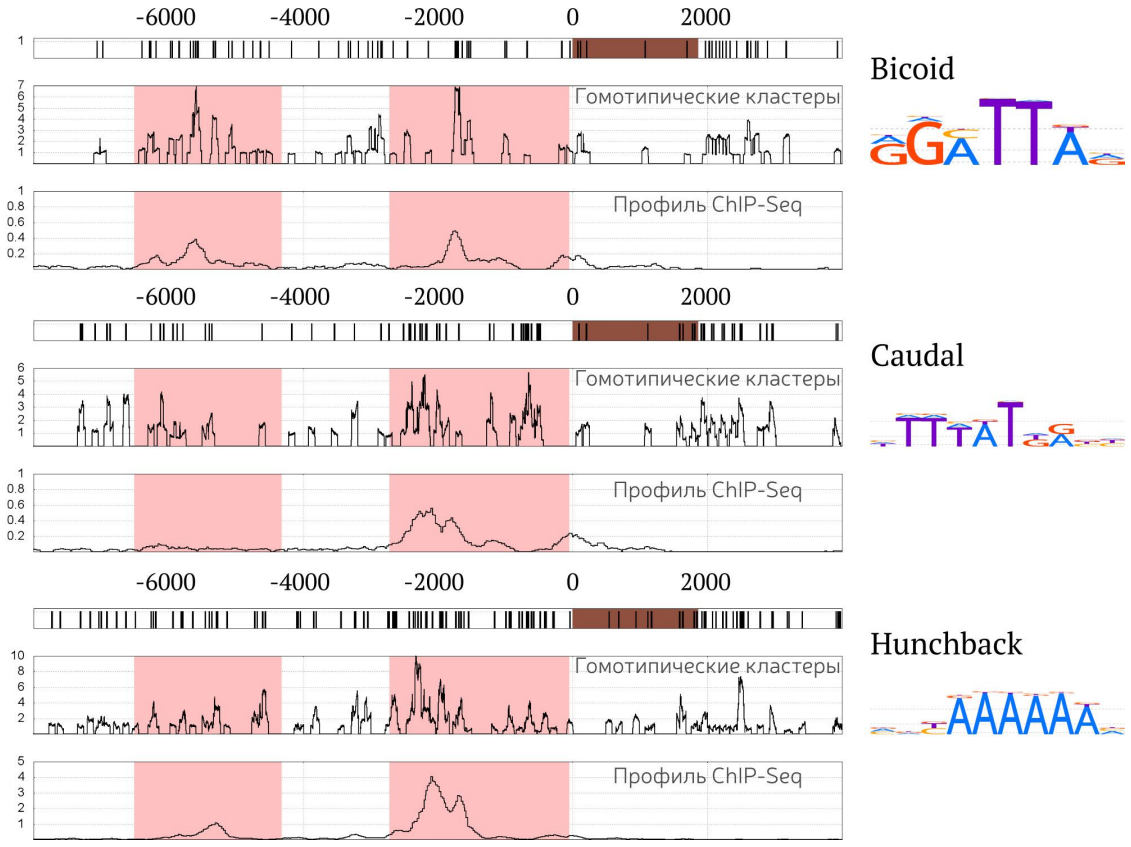


Рисунок 14. Визуализация экспериментальной (ДНКазо-доступность, ChIP-Seq) и вычислительной (мотивы) аннотации локуса гена *giant* в геноме *Drosophila melanogaster*.

Показаны группы треков для трех факторов: (*верхняя панель*) Bicoid, (*средняя панель*) Caudal, (*нижняя панель*) Hunchback. Треки в каждой группе: (*верхний трек*) предсказанные сайты связывания, темным фоном выделен кодирующий район гена; (*средний трек*) предсказанные гомотипические кластеры сайтов связывания, темным фоном выделены ДНКазо-доступные районы; (*нижний трек*) пики ChIP-Seq, темным фоном выделены ДНКазо-доступные районы. Ось X соответствует локальной геномной координате (0 – аннотированный старт транскрипции гена), ось Y показывает статистическую значимость (для гомотипических кластеров) и высоту пика (для ChIP-Seq). Экспериментальные данные соответствуют 5 стадии развития эмбриона. Рисунок адаптирован из обзора [Kulakovskiy, Makeev, 2013].

3.3.2.6. Систематические ошибки ChIP-Seq

...enrichment in a given ChIP experiment depends on many intractable parameters, likely including a phase of a moon.

… обогащение (прочтениями) в конкретном ChIP-эксперименте зависит от массы трудноотслеживаемых параметров, вероятно, включая фазу Луны.

[Barski, Zhao, 2009]

Фактически ChIP-Seq сделал возможной масштабную количественную оценку связываемых белком сегментов генома *in vivo* вплоть до однонуклеотидного разрешения, при условии тщательного постпроцессинга. При всех формальных и реальных преимуществах технологии использование и обработка данных ChIP-Seq сталкивается с рядом практических проблем.

Во-первых, результаты ChIP-Seq зависят от качества антител, последствия использования различных антител могут быть диаметральными – от чрезвычайно сильных пиков, соответствующих нерелевантным белкам, до полного отсутствия наблюдаемой кластеризации чтений.

Во-вторых, использование соникации для фрагментации ДНК порождает так называемый эффект «Sono-Seq»: эффективность фрагментации хроматина положительно скоррелирована с транскрипционной активностью промоторов и доступностью хроматина в различных районах хромосом [Auerbach и др., 2009; Jain и др., 2015]. Это может вызывать ошибочное обнаружение связывания удивительно нерелевантных белков, например комплексов сайленсинга (выключения транскрипции) непосредственно в наиболее активно транскрибируемых районах [Teytelman и др., 2013].

Артефакты, вызванные систематическим вкладом нуклеосомной укладки в процесс фрагментации и экстракции ДНК, могут давать ошибочные глобальные наблюдения о совместной локализации или избегании нуклеосом конкретными факторами транскрипции [Park и др., 2013].

Районы, избыточно часто детектируемые в различных несвязанных экспериментах, можно исключить с использованием «черных списков»[20], например, построенных участниками проекта ENCODE [Dunham и др., 2012]. Артефакты,

[20] Blacklisted genomic regions for functional genomics analysis. https://sites.google.com/site/anshulkundaje/projects/blacklists

связанные с качеством антител и доступностью хроматина, можно побороть с использованием двойного контроля (секвенирования и геномной ДНК, и результатов неспецифичной иммунопреципитации). Трудоемкий способ – добиваться стабильных экспериментальных повторностей и использовать оценку невоспроизводимости в биологических повторностях (irreproducible discovery rate, IDR) как основной критерий достоверности пиков [Landt и др., 2012]. Альтернативный и более надежный подход для избавления от нерелевантных пиков – нокаутирование целевого гена [Krebs и др., 2014] и дифференциальный анализ результатов для дикого типа/нокаута.

Практическая применимость методов удаления артефактов (в том числе, возможность дедупликации чтений в «бедных» библиотеках секвенирования) подробнее обсуждается в работе [Carroll и др., 2014].

Решающим аргументом может быть анализ мотивов, который предоставляет важную информацию о качестве данных (локализация мотива в пиках, обогащенность вхождениями) и, до определенной степени, по наличию вхождений мотива позволяет разделить события прямого и непрямого (piggy-back, через партнера) связывания ДНК.

3.3.2.7. Идентификация мотивов в ChIP-Seq данных

Для использования мотивов в анализе и контроле качества ChIP-Seq данных необходимо, собственно, откуда-то получить модель мотива. Для хорошо изученных факторов транскрипции модель мотива может быть уже известна, например, построена ранее на основе результатов классических «догеномных» экспериментальных методов. Но как быть с результатами экспериментов для слабоизученных белков?

В момент появления высокопроизводительных методов, в том числе, методов на основе иммунопреципитации хроматина, стало понятно, что существующие инструменты идентификации мотивов не в состоянии справиться с обработкой тысяч и десятков тысяч длинных последовательностей. Это связано как с объемом данных, так и с ограниченной разрешающей способностью высокопроизводительных экспериментов. Поиск короткого сигнала ДНК-белкового узнавания из 10-20 букв в последовательностях длиной в сотни и тысячи нуклеотидных оснований имеет

опасное сходство с поиском иголки в стоге сена. Первые успешные инструменты, способные управляться с данными ChIP-chip, использовали строковые методы идентификации мотивов, основанные на выявлении перепредставленных подстрок с помощью суффиксных деревьев [Ettwiller и др., 2007] и подсчета *k*-меров [Linhart, Halperin, Shamir, 2008]; и только на последних стадиях алгоритмов кандидатные подстроки превращались в весовые матрицы. Такая процедура позволяет значительно сократить вычислительное время и выявить не только основной мотив изучаемого белка, но и мотивы кофакторов. Очевидный минус состоит в том, что получаемая из анализа подстрок и *k*-меров модель в форме весовой матрицы не является оптимальной в смысле точности описания паттерна сайтов связывания, то есть ограничена применима для распознавания сайтов – последующего поиска наилучших вхождений мотивов.

Альтернативный подход, позволяющий использовать весовые матрицы сразу на этапе оптимизации и ограничить пространство поиска для экономии вычислительных ресурсов, – использование априорной информации. Для ChIP-chip в этой роли можно было привлекать данные догеномных методов. Так, прямое включение догеномных экспериментальных данных на этапе построения множественного выравнивания позволяло успешно получать из ChIP-chip достаточно точные весовые матрицы [Kulakovskiy, Makeev, 2009]. Тем не менее, по-настоящему устойчивого и универсального инструмента для построения точной модели мотива только по данным ChIP-chip предложено не было. К счастью для анализа мотивов, ChIP-chip быстро уступил свое место технологии ChIP-Seq, которая, как мы уже говорили, не только обладает лучшим базовым разрешением, но и содержит сопутствующую информацию о форме пиков.

Для идентификации мотива заданного фактора транскрипции *de novo* естественным кажется использование полного набора пиков ChIP-Seq, определеных по результатам конкретного эксперимента. В то же время, различные поднаборы пиков могут содержать различные подтипы мотивов [Levitsky и др., 2016], и априорно неизвестно как выбрать оптимальный набор пиков, чтобы обеспечить хорошо сэмплированное множество сайтов связывания. Например, более высокие пики потенциально содержат более сильные сайты связывания, а более широкие пики – кластеры средних по аффинности сайтов связывания. В реальности задача

оказывается еще сложнее. Экстенсивный путь – использовать полный набор пиков – может быть принципиально ошибочным: если слабые пики присутствуют в большом количестве и соответствуют слабым сайтам связывания, то они будет вносить существенный шум в определение слабо выраженных позиций сайтов связывания, например, позиций фланкирующих коровый участок мотива.

Вопреки значительной неопределенности, для ChIP-Seq довольно быстро появились разумные эвристические соображения, которые позволили применять традиционные программы, в том числе MEME [Bailey и др., 2009], для идентификации мотивов [Guillon и др., 2009; Jothi и др., 2008; Tuteja и др., 2009; Valouev и др., 2008]. Стандартный подход состоит в том, чтобы для идентификации мотива взять только небольшой поднабор наилучших пиков (топ-100, топ-500). По крайне мере, такой набор по построению содержит меньше ошибочно определенных регионов связывания (быть может, за исключением экстремально высоких артефактных пиков), и в среднем соответствующие пики содержат сайты связывания с высокой аффинностью, что позволяет идентифицировать устойчивый мотив хотя бы для наиболее выраженных сайтов.

Еще одна эвристика эксплуатирует наблюдение о колоколообразной форме пиков. Фланкирующие регионы у таких пиков могут быть безопасно обрезаны с сохранением региона в районе вершины. К сожалению, реально при обрезке редко оценивают ширину колокола, а чаще используют фиксированные 300-500 нуклеотидов вокруг вершины, для пиков сложной формы часть сайтов связывания может быть утеряна, но это частично компенсируется эффектом кластеризации.

Две описанных эвристики оказываются достаточными, чтобы использовать «догеномные» методы идентификации мотивов для анализа ChIP-Seq данных. В то же время, не хочется на этом этапе отказываться от подробной информации о форме пиков, или хотя бы о локализации вершины; тем более что существуют возможности для явного включения априорной информации в идентификацию мотивов [Bailey и др., 2010; Simcha, Price, Geman, 2012].

Судя по всему, первой программой, специально созданной для использования позиционной информации о профиле покрытия пиков фрагментами ДНК, была HMS (Hybrid Motif Sampler [Hu и др., 2010]), модификация классического Гиббсовского сэмплера [Neuwald, Liu, Lawrence, 1995], впоследствие даже портированная для

вычисления на графических процессорах GPGPU [Zandevakili, Hu, Qin, 2012]. К сожалению, использование профиля покрытия не всегда помогало HMS в распознавании правильного мотива связывания. Практическую успешность идея использования профиля покрытия приобрела после публикации ChIPMunk [Kulakovskiy и др., 2010], модификации жадного алгоритма для поиска мотивов.

Говоря о мотивах в ChIP-Seq, хочется отметить два замечательных факта. Во-первых, использование ChIP-Seq позволило значительно улучшить классические модели мотивов в форме позиционно-весовых матриц, и средняя точность весовых матриц, основанных на ChIP-Seq данных [Kulakovskiy и др., 2013a; Kulakovskiy и др., 2016; Wang и др., 2012b; Wang и др., 2013a], стабильно и заметно превосходит мотивы из ранних коллекций [Dabrowski и др., 2015; Kibet, Machanick, 2015]. Во-вторых, стремительное накопление данных ChIP-Seq для широкого спектра факторов транскрипции открыло возможности и для систематического построения расширенных моделей сайтов связывания, и для масштабного тестирования моделей с точки зрения качества распознавания сайтов связывания.

Расширенные модели сайтов связывания могут учитывать корреляции между соседними либо удаленными нуклеотидами, особенности нуклеотидного состава в позициях, фланкирующих кор мотива и присутствие подтипов мотива (например, подмотивов различной длины состоящих из отдельных боксов с переменным спейсером). Расширение модели может быть достигнуто различными методами: используя смесь весовых матриц (выбор конкретной матрицы из смеси определяется нуклеотидами в ключевых позициях сайтов связывания [Bi и др., 2011]), путем явного подсчета динуклеотидных весов [Kulakovskiy и др., 2013b; Levitsky и др., 2014], с использованием скрытых марковских моделей [Mathelier, Wasserman, 2013], байесовских марковских моделей переменного порядка [Siebert, Söding, 2016] или обобщенных схем, назначающих веса и близким и далеким зависимостям между позициями [Keilwagen, Grau, 2015]. Практический инструментарий для анализа мотивов с использованием расширенных моделей пока не так хорошо развит, как для ПВМ. Каждый новый подход претендует на роль золотого стандарта и не очевидно, кто выйдет победителем: дело не только и не столько в самой математической модели, сколько в методике ее обучения по экспериментальным данным (например, ряд упомянутых подходов при обучении модели может опираться на стартовые

выравнивания, полученные с помощью ПВМ). Можно ожидать, что наилучший подход в обозримом будущем будет определен в ходе независимых тестирований и кроссвалидации на различных экспериментальных данных.

Авторская программа ChIPMunk, ее расширение для учета корреляций между соседними нуклеотидами и созданная коллекция моделей сайтов связывания для факторов транскрипции мыши и человека, подробно обсуждаются ниже в разделах «Материалы и методы» и «Результаты и обсуждение».

3.3.2.8. Программные инструменты и практический анализ ChIP-Seq данных

Программное обеспечение в активно развивающихся областях биоинформатики, к сожалению, не всегда стабильно развивается и поддерживается; а рекомендации производителя могут отсылать не только к фирменному коммерческому программному обеспечению, но и обратно к исследовательским обзорам. Поддерживаемые сообществом сайты-каталоги программ[21,22] часто оказываются существенно актуальнее, чем научные статьи, описывающие устаревающие или уже забытые подходы. Тем не менее, отметим ряд базовых обзоров и полезных программ, достаточных для анализа данных ChIP-Seq от прочтений до аннотированных сайтов связывания.

По вопросу о методах картирования чтений стоит обратиться к систематическим обзорам [Flicek, Birney, 2009; Li, Homer, 2010] и учесть, что массово применяемые инструменты, такие как Bowtie [Langmead и др., 2009] и BWA [Li, Durbin, 2009], пусть и не обладают наилучшими возможными показателями точности и чувствительности картирований, но дают разумный баланс между производительностью и качеством картирования; а кроме того имеют широкую базу пользователей и хорошо поддерживают стандартные форматы ввода-вывода. Принципиальные моменты, на которые стоит обратить внимание при выборе параметров картирования для ChIP-Seq: стоит избегать неравномерности по цепям (strand bias) и осторожно ограничивать неуникальные мультикартирования (в том числе, соответствующие геномным повторам). И то и другое может негативно сказываться на идентификации пиков или последующем анализе мотивов

[21] SEQanswers wiki. http://seqanswers.com/wiki/SEQanswers
[22] OMICStools. https://omictools.com

связывания. Уже разработаны подходы для сохранения неуникальных картирований [Chung и др., 2011; Wang и др., 2013b], но пока неясно, насколько это продуктивно в общем случае.

С выбором инструментов для идентификации пиков все также непросто. Для базовой оценки обогащения конкретного сегмента генома используется широкий спектр подходов: от простого подсчета перекрывающихся фрагментов ДНК, полученных расширением чтений [Fejes и др., 2008], до моделирования сдвига между картирования чтений по разным цепям [Zhang и др., 2008] и байесовской сегментации генома на обогащенные и необогащенные прочтениями участки [Xing и др., 2012]. На стартовом этапе заслуженной популярностью пользовались сравнительно простая программа FindPeaks [Fejes и др., 2008] и сопутствующий программный пакет для базового анализа данных секвенирования[23]; к сожалению проект лишился поддержки и документации с уходом автора из академической науки в индустрию.

На сегодня стандартом является метод сравнения локального обогащения в эксперименте с контролем: для оценки вероятности пронаблюдать заданное число чтений в локальном окне используется биномиальный тест или распределение Пуассона [Thomas и др., 2016], параметры которого оцениваются по контрольным данным.

Так или иначе, итоговые пики, определенные различными программами, заметно отличаются и по длине, и по общему числу. Детальные сравнительные обзоры и тестирования программ, безусловно, публиковались [Kim и др., 2011; Rye, Sætrom, Drabløs, 2011; Wilbanks, Facciotti, 2010], но постоянное появление новых программ стремительно опережает усилия по сравнительному тестированию существующих инструментов [Szalkowski, Schmid, 2011]. Попытка совместного тестирования попарных комбинаций программ [Schweikert и др., 2012] показала разумный ожидаемый эффект: объединение результатов разных программ улучшает чувствительность, а пересечение – специфичность; но тестирование не позволило сформулировать четкой стратегии «как выбрать программу или какую комбинацию программ в общем случае следует использовать в качестве базового варианта».

[23] Vancouver Short Read Analysis Package. http://vancouvershortr.sourceforge.net

Сегодня для идентификации пиков в онлайн-каталоге OMICtools можно найти уже несколько десятков инструментов и их число продолжает расти. В то же время, провал в сравнительном тестировании сохраняется; так, в относительно поздней работе [Thomas и др., 2016] сделан детальный разбор отличий и возможных преимуществ используемых статистических моделей, но на практике протестировано только 5 вычислитеьных методов и только на данных 3 экспериментов (из которых два эксперимента – по анализу гистоновых меток и только один – для фактора транскрипции); а итоговые выводы о «полезности» или «применимости» алгоритмов даются в очень осторожном ключе.

Тот факт, что задача выбора подходящей программы для поиска пиков в ChIP-Seq данных не решена, отмечают и другие авторы [Nakato, Shirahige, 2016]. Отдельная проблема состоит в том, что не все старые программы способны корректно обрабатывать выравнивания прочтений в современных форматах и тестирование сталкивается не только с содержательными, но и с техническими проблемами.

Основываясь на различных обзорах и практическом опыте, мы можем дать две общие рекомендации. Во-первых, в качестве базовой программы для идентификации пиков можно смело использовать MACS [Feng и др., 2012]. Эта программа зарекомендовала себя пусть не наилучшим, но стабильным поведением в различных сравнительных тестированиях; а кроме того, продолжает активно развиваться и поддерживаться с момента исходной публикации [Zhang и др., 2008] вплоть до текущего времени (2017). В случае, когда весь анализ ведется в статистическом окружении R с пакетом Bioconductor для поиска пиков можно предпочесть PICS [Mercier и др., 2011] или SPP [Kharchenko, Tolstorukov, Park, 2008], чтобы проводить весь анализ средствами R. ChIPpeakAnno – еще один полезный инструмент, представленный в R/Bioconductor – проводит аннотацию пиков по близлежащим генам и другим элементам разметки генома [Zhu и др., 2010].

Стоит отметить, что большинство открытых программ сегодня используют интерфейсы командной строки, инструменты с выделенным графическим интерфейсом (за пределами коммерческих пакетов[24]) не получили широкого распространения, например, популярный некогда CisGenome [Ji и др., 2008]

[24] CLC Genomics Workbench. http://www.clcbio.com

достаточно давно не обновляется. Впрочем, появляются новые сервисы, призванные предоставить исследователям-биологам визуальный интерфейс для анализа данных ChIP-Seq; например, интегрированная среда EaSeq [Lerdrup и др., 2016]. Можно только приветствовать разработку наглядной визуализации и интерактивных инструментов для анализа данных, не требующих навыков работы в командной строке, но пока, в отсутствие независимого содержательного тестирования, трудно сказать что-то определенное о достоверности получаемых результатов. Это верно, к сожалению, и для коммерческих программных пакетов.

Существуют интересные попытки публикации готовых «конвейеров обработки» из стандартных программ для запуска и инсталляции «одним кликом» на персональном компьютере, например «Fish the ChIPs» [Barozzi и др., 2011] для системы MacOS X; но неочевидно как это поможет пользователям других операционных систем и насколько успешно такой «готовый» конвейер справляется с обновлением программ в своем составе.

Именно поэтому биологам, сталкивающимся с необходимостью самостоятельного анализа данных и не имеющим достаточных компьютерных навыков для работы в командной строке или в окружении R, можно в первую очередь рекомендовать использовать публичные серверы Galaxy, например, Nebula [Boeva и др., 2012], где необходимые стандартные и хорошо зарекомендовавшие себя инструменты собраны в единый интерфейс, пусть и лишенный гламурного лоска.

В контексте анализа мотивов нельзя не упомянуть группу программ для идентификации пиков, совмещенных с одновременным поиском мотива связывания: сюда входит MICSA [Boeva и др., 2010] (интеграция FindPeaks + MEME) и GEM [Guo, Mahony, Gifford, 2012]. Что касается отдельного анализа мотивов, биологам могут оказаться полезными peak-motifs в составе RSAT [Thomas-Chollier и др., 2012] и MEME-ChIP [Bailey, Machanick, 2012] в составе MEME suite [Bailey и др., 2009]. Web-интерфейсы в среднем не предоставляют богатых возможностей для специфического кастомизированного анализа, но базовую информацию о присутствии и локализации мотивов получить можно; обзор имеющихся в сети инструментов сделан в работе [Tran, Huang, 2014]. Для изучения колокализации или позиционных предпочтений мотивов для кофакторов изучаемого белка можно

применить rGADEM [Mercier и др., 2011], SpaMo [Whitington и др., 2011] или COPS [Ha, Polychronidou, Lohmann, 2012].

Несмотря на все разнообразие инструментов, часть технических задач все еще приходится решать в индивидуальном порядке, то есть с помощью простейших средств автоматизации или скриптов. Это вызвано отсутствием общепринятых стандартов на форматы ввода-вывода для программ идентификации мотивов и поиска мотивами и необходимостью совмещения информации о нуклеотидных последовательностях, соответствующих пикам, и профилей покрытия чтениями. Что касается стандартных форматов файлов геномных разметок (в том числе пиков), хорошая документация присутствует в справочной информации геномного браузера UCSC Genome Browser[25] [Fujita и др., 2011].

Мы считаем, что систематический подход в анализе ChIP-Seq должен заключаться в последовательном пошаговом использовании наиболее стабильных инструментов для картирования, идентификации пиков, анализа мотивов с контролем результатов на каждой стадии. В этом смысле, наиболее логично на наш взгляд выглядит простой конвейер обработки [Maksimenko и др., 2015], составленный из Bowtie [Langmead, Salzberg, 2012] и MACS [Feng и др., 2012] для картирования прочтений и идентификации пиков, ChIPMunk [Kulakovskiy и др., 2010; Kulakovskiy и др., 2013b] и SARUS [Kulakovskiy и др., 2016] для анализа мотивов, с последующей аннотацией в R: привязкой к генам с помощью ChIPpeakAnno [Zhu и др., 2010] и статистической оценкой колокализации с «геномными разметками» (например, с пиками ChIP-Seq других белков) с помощью GenometriCorr [Favorov и др., 2012].

3.3.2.9. Дальнейшая эволюция ChIP-Seq для факторов транскрипции

Пик ChIP-Seq получается в результате сложения сигналов от относительно протяженных фрагментов ДНК, перекрывающихся с сайтом или сайтами связывания целевого белка. Естественное желание – уменьшить реальную длину фрагментов ДНК, чтобы улучшить разрешение. Изящное решение под названием ChIP-exo было предложено в работах [Rhee, Pugh, 2011; Rhee, Pugh, 2012], а именно, иммунопреципитированные фрагменты ДНК дополнительно обрабатываются λ-

[25] UCSC Genome Bioinformatics FAQ. https://genome.ucsc.edu/FAQ/FAQformat.html

экзонуклеазой, в результате прочитываемые 5’-концы фрагментов уже практически точно соответствуют положению сайта связывания в геноме. Улучшение в точности позиционирования сайтов связывания сравнимо с переходом от гибридизации к секвенированию, см. **Рисунок 15**. Кроме того, неспецифически иммунопреципитированная ДНК (лишенная сшивок с исследуемым белком) деградируется экзонуклеазой, что теоретически должно уменьшать фоновый шум.

По всей видимости, подходы в идентификации пиков и мотивов для ChIP-exo переносимы с ChIP-Seq напрямую или с минимальными модификациями. Тем не менее пока неясно сможет ли ChIP-exo приобрести большую популярность и насколько воспроизводимыми окажутся результаты в случае слабых антител или сильно перекрывающихся сайтов связывания (в этом смысле ChIP-Seq выглядит более устойчивым и в смысле точности геномного картирования длинных прочтений и в смысле усреднения и сглаживания локальных флуктуаций сигнала). Также неясно, что может быть заменой input-контроля для ChIP-exo: очищенная геномная ДНК не защищена от деградации экзонуклеазой и, следовательно, не может использоваться напрямую.

Технология ChIP-Seq даже в узкой нише анализа сайтов связывания факторов транскрипции продолжает получать новые интересные приложения, например, можно проводить анализ аллель-специфичного связывания путем рассмотрения полиморфных позиций в выравниваниях чтений [Cavalli и др., 2016; Ni и др., 2012]. В то же время, несмотря на достаточно большую историю успешного и разнообразного использования, все еще не закрыта масса вопросов о воспроизводимости и фильтрации артефактных сигналов, связанных с неравномерной доступностью хроматина. Технология высокопроизводительного анализа транскриптома, RNA-Seq, является гораздо более зрелой и технически более простой (хотя бы потому, что размер транскриптома для многих организмов хорошо известен, нет «проблемы антител» и необходимости дополнительных контролей). Но даже для RNA-Seq работы по действительно массовому анализу стабильности результатов в экспериментальных повторностях и внутри популяций начали появляться лишь недавно [Lin и др., 2016; Schurch и др., 2016]. Перспективы проведения подобных крупномасштабных исследований для связывания белков с ДНК выглядят интригующе, но туманно.

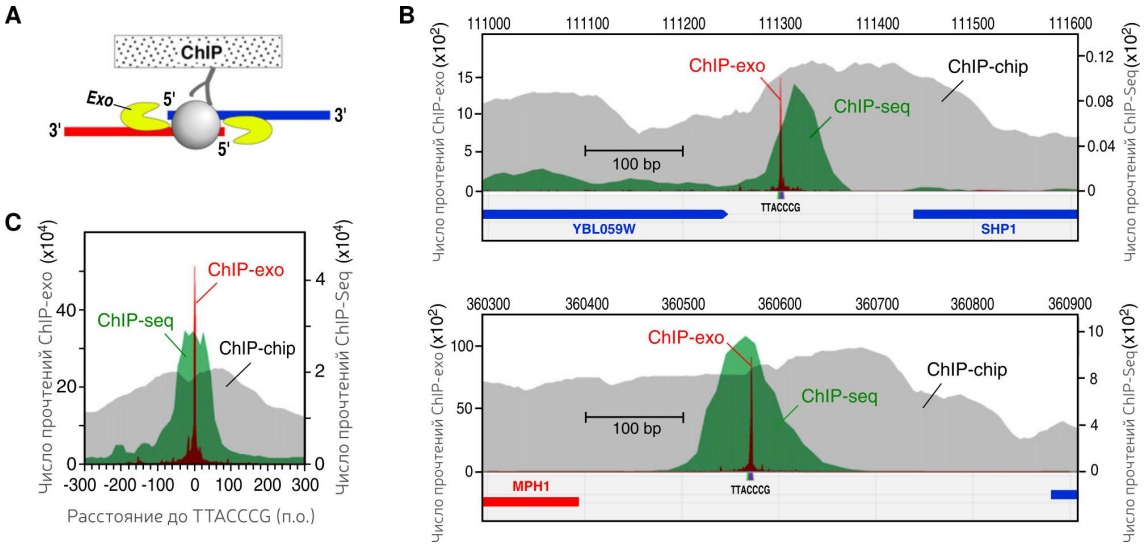


Рисунок 15. (*A*) Идея метода ChIP-exo. (*B*) Сравнение разрешения ChIP-exo, ChIP-chip и ChIP-Seq для фактора транскрипции Reb1 в конкретном локусе генома дрожжей. (*C*) Усредненный сигнал Reb1 для 791 вхождений TTACCCG в геноме дрожжей.

Рисунок адаптирован из работы [Rhee, Pugh, 2011].

3.3.3. Сложность интерпретации результатов высокопроизводительных экспериментов

Данные о том, что связывание факторов транскрипции может не оказывать прямого влияния на экспрессию близлежащих генов, появились достаточно давно [Fernandez и др., 2003]. Более того, с развитием высокопроизводительных методов стало все труднее игнорировать массовое присутствие выраженных сайтов связывания в межгенных районах, в значительном удалении от кодирующих областей [Jothi и др., 2008; MacQuarrie и др., 2011]. Долгое время достоверность таких наблюдений ставилась под сомнение из-за особенностей экспериментов по иммунопреципитации хроматина. Критику нельзя полностью игнорировать: массовый анализ экспериментальных данных ChIP-Seq, имеющихся в открытом доступе, выявил отрицательную корреляцию между импакт-фактором журнала и оценкой качества ChIP-Seq данных с точки зрения кластеризации чтений [Marinov и др., 2014]. Можно предположить, что высокоимпактные статьи часто посвящены слабоизученным и/или сложным для анализа фактором транскрипции (для которых, например, отсутствуют коммерческие антитела и нетривиальным может быть их самостоятельное получение) и массовым экспериментам, а именно массовые эксперименты особенно страдают от отсутствия тщательного контроля качества и невозможности «ручного» анализа данных на каждой стадии. Есть и позитивный факт: в ходе наиболее крупных международных проектов по аннотации регуляторных районов генома контроль за анализом данных проводится достаточно тщательно.

Формально, крупномасштабные проекты преследовали амбициозные цели: с помощью современных высокопроизводительных методов «избавиться от белых пятен на карте генома» и получить подробную функциональную аннотацию для различных типов клеток. На наш взгляд, наиболее ценным реальным результатом стал репозиторий экспериментальных данных с контролируемой достоверностью и стандартизация сопутствующих экспериментальных протоколов и методов анализа.

Наиболее известный крупномасштабный проект – ENCODE[26] (Encyclopedia of DNA Elements) [Myers и др., 2011] – был анонсирован еще в 2003 году, а в 2007 опубликовал результаты пилотного исследования [Birney и др., 2007] и приступил к

[26] ENCODE: Encyclopedia of DNA Elements. https://www.encodeproject.org/

фазе массовых экспериментов. В ходе работы планировалось проаннотировать всевозможные функциональные районы и элементы генома эукариот: генетические и эпигенетические, от количественных данных по транскриптомам до сайтов связывания факторов транскрипции (в первую очередь проект фокусировался на человеке и мыши; модельные организмы *D. melanogaster* и *C. elegans* изучались сестринским проектом modENCODE). С каждым годом рос и охват типов клеток (от иммортализованных клеточных линий к первичным клеткам и образцам тканей) и типов экспериментальных данных (так в 2016 году были открыты для публичного доступа данные Hi-C [Berkum van и др., 2010] по картами хромосомных контактов).

Дэн Гроур (Dan Graur), эволюционный биолог, скандально известный за едкие публикации как в научных журналах, так и в личном блоге, издевательски сравнивал[27] тон публикаций ENCODE 2012 и 2014 года [Dunham и др., 2012; Muerdter, Stark, 2014], в которых уверенный оптимизм сменился выверенной осторожностью формулировок. С последовательным ростом объема данных пришло осознание, что серийные эксперименты не смогли выявить функцию некодирующих районов ДНК (и, в том числе, объяснить роль связывания факторов транскрипции, не ассоциированную с регуляцией экспрессии): “*The papers do not reveal how many of these elements might be functional, and independent estimates span a broad range*” («Публикации не проливают свет на то, сколько из обнаруженных элементов являются функциональными, а независимые оценки существенно отличаются» [Muerdter, Stark, 2014]).

Тем не менее, нельзя недооценивать практическую ценность результатов проекта ENCODE, который уже опубликовал в открытом доступе широчайший спектр единообразно собранных и обработанных экспериментальных данных. В ходе проекта проводится важная работа по оптимизации протоколов для массового серийного использования и созданию практических рекомендаций по организации экспериментов и компьютерной обработке данных (например, для ChIP-Seq [Landt и др., 2012]). Число научных публикаций, использующих данные ENCODE, исчисляется уже тысячами, и их число продолжит расти; активно использует результаты ENCODE и эта диссертационная работа.

---

[27] ENCODE 2014 versus ENCODE 2012 (with translations) @ENCODE_NIH. http://judgestarling.tumblr.com/post/95976986801/encode-2014-versus-encode-2012-with-translations

3.4. Перспективные приложения мотивов

Регуляция транскрипции у эукариот – сложный процесс, и в смысле разнообразия элементов регуляторного механизма, и в смысле методов его изучения. Экспериментальные методы постепенно эволюционируют как количественно (увеличивается объем получаемых данных, улучшается разрешение), так и качественно. Получение карты сайтов связывания, специфичной для типа клеток или условий, не видится самодостаточным и даже массовые эксперименты постепенно двигаются в сторону содержательных вопросов: является ли активная транскрипция маркером реальной промоторной или энхансерной области, является ли связывание регуляторного белка определяющим для экспрессии конкретных генов, интересных для фундаментальной науки или медицинских приложений.

Возможности для приложения методов анализа мотивов, с другой стороны, только расширяются. Появляются данные о прямой роли промоторов и факторов транскрипции в определении дальнейшей судьбы мРНК: стабильности [Talarek, Bontron, Virgilio De, 2013; Trcek и др., 2011], локализации и трансляции [Zid, O'Shea, 2014]. Пока этот феномен активно изучается у дрожжей, но, судя по всему, похожие механизмы есть и у высших эукариот [Dori-Bachash и др., 2012]. То есть, аннотация промоторов становится важной не только для уровня транскрипции, но и для предсказания пост-транскрипционной регуляции. С другой стороны, применение мотивов не ограничивается факторами транскрипции. Появляются прямые экспериментальные данные о паттернах, распознаваемых РНК-связывающими белками [Ray и др., 2013], что позволяет использовать мотивы для аннотации последовательностей РНК. Характерные паттерны-мотивы выделяются и в других случаях, например, для контекстов мутаций, вносимых дезаминазами семейства APOBEC [Lada и др., 2013], играющих важную роль в соматическом мутагенезе в раковых клетках [Alexandrov и др., 2013]. Наконец, с помощью анализа мотивов удается частично перейти и на уровень эпигенетики, пока в виде эпигенетической карты генома [Quante, Bird, 2016; Whitaker, Chen, Wang, 2015]. Все это еще раз подчеркивает важность роли нуклеотидных мотивов и разнообразие функций, которые реализуются через них, в ходе направленной реализации генетической информации.

4. Материалы и методы

В этом разделе диссертации в деталях рассматриваются авторские вычислительные методы анализа мотивов: идентификация мотивов в результатах экспериментов, основанных на техниках высокопроизводительного секвенирования, сравнение мотивов и аннотация регуляторных вариантов в последовательностях. Специальные особенности и параметры использованных подходов затем дополнительно обсуждаются в конкретных секциях раздела «Результаты и обсуждение».

4.1. Идентификация мотивов в больших выборках нуклеотидных последовательностей. Алгоритм ChIPMunk

4.1.1. Мотивация разработки алгоритма

Замена гибридизации на микрочипах секвенированием (переход от ChIP-chip к ChIP-Seq) снабдила геномные данные по взаимодействиям белок-ДНК детальным позиционным профилем, характеризующим связывание регулятора в конкретной области. Полноценный метод идентификации мотивов сайтов ДНК-белкового узнавания должен учитывать априорное знание о предполагаемой локализации сайтов в относительно протяженных пиках ChIP-Seq – «форму» пиков, т.е. профиль покрытия сегментов прочтениями. И кроме того, необходимо обеспечить обработку больших выборок данных.

Первым методом, явно эксплуатирующим форму пиков, по всей видимости, был HMS [Hu и др., 2010], его практическое применение было ограничено нестабильными результатами и сложным форматом входных данных. Мы, в свою очередь, предложили модификацию жадного алгоритма ChIPMunk [Kulakovskiy и др., 2010], которая показала хорошие результаты как в нашем сравнительном тестировании, так и в последующих независимых бенчмарках [Kuttippurathu и др., 2011; Ma и др., 2012].

Исходная версия алгоритма ChIPMunk, опубликованная в 2009 году [Kulakovskiy, Makeev, 2009], была предложена для поиска единого мотива в разнородных экспериментальных данных (включая «большие» данные ChIP-chip). В обновленной версии [Kulakovskiy и др., 2010], помимо учета позиционных профилей

ChIP-Seq, значительно расширена базовая функциональность. Ключевые идеи и их реализация обсуждаются ниже.

4.1.2. Ключевые идеи и формализация

Словарные методы, основанные на различных методах подсчета частот олигонуклеотидов – «слов» на 4-х буквенном алфавите {A,C,G,T}, справляются с большими наборами последовательностей значительно быстрее, чем методы с использованием обобщенных представлений мотивов, даже в сравнительно простой форме позиционно-весовых матриц. В то же время, получение относительно точного описания мотива имеет самостоятельную ценность, в том числе, для предсказания сайтов связывания в протяженных последовательностях, или для корректного ранжирования сайтов по аффинности связывания и дизайна минимальных мутаций, необходимых и достаточных для изменения аффинности (например, для последующего экспериментального тестирования функциональности сайтов связывания в зависимости от геномных вариантов).

Алгоритм ChIPMunk нацелен на построение наиболее точной ПВМ и последовательно решает две задачи: (а) построение псевдо-оптимального безделеционного локального выравнивания последовательностей (gapless multiple local alignment), определяющего наилучшее (наиболее похожее на мотив) слово в каждой последовательности и (б) подбор порогового значения оценки весовой матрицы для сегментации выравнивания на «сайты» и «не-сайты» на основе самосогласованности весовой матрицы. По умолчанию ChIPMunk рассматривает обе цепи каждой последовательности, поскольку алгоритм исходно проектировался для мотивов сайтов связывания факторов транскрипции. В поздних версиях добавлен одноцепочечный режим, подходящий для анализа мотивов РНК-связывающих белков.

Получаемые ChIPMunk выравнивания являются субоптимальными в силу стохастического алгоритма, основанного на случайном сэмплировании обучающей выборки. В то же время, поиск действительно оптимального выравнивания не обязателен, поскольку обучающая выборка всегда является каким-то подмножеством «истинных возможных» сайтов связывания, и ее наилучшее выравнивание не является действительно наилучшей моделью сайтов связывания. То есть, среди

множества возможных решений достаточно выбрать любое из подмножества «достаточно хороших», способных сравнимо точным образом описать множество сайтов связывания в учебной выборке и в независимых контрольных данных.

4.1.2.1. Оптимальность множественного локального выравнивания последовательностей. Дискретное информационное содержание с учетом расстояния Кульбака-Лейблера

Среди множества возможных выравниваний ChIPMunk стремится выбрать мотив с наилучшим дискретным информационным содержанием, т.е. с наиболее выраженным преобладанием конкретных нуклеотидов в конкретных позициях. Идея оптимизации информационного содержания при построении мотива не нова [Hertz, Stormo, 1999]. Мы используем дискретное информационное содержание (ДИС, discrete information content, DIC) с кульбаковским членом (КДИС, KDIC).

Обозначим распределение фоновых частот нуклеотидов в изучаемых последовательностях как $Q = \{q_1, .., q_k\}$, где $\{1, .., k\} = \{A, C, G, T\}$. Рассмотрим колонку j в выравнивании N последовательностей; пусть частоты нуклеотидов (букв) в этой колонке равны $\{x_1, .., x_k\}$: $\sum_{i=1}^{k} x_i = N$. Вероятность получить фиксированные частоты нуклеотидов как в заданной колонке j множественного выравнивания N последовательностей – на основе серии из N независимых испытаний при заданных вероятностях $Q = \{q_1, .., q_k\}$ равна:

$$P(x_1, .., x_k | Q, j) = \frac{N!}{x_1! .. x_k!} q_1^{x_1} .. q_k^{x_k} \quad (1).$$

Взяв логарифм получим:

$$\log P(x_1, .., x_k | Q, j) = \log \frac{N!}{x_1! .. x_k!} + \sum_{i=1}^{k} x_i \log q_i \quad (2).$$

Ранее в работе [Kulakovskiy, Favorov, Makeev, 2009] мы ввели дискретное информационное содержание (ДИС) как меру гомогенности распределения нуклеотидов в колонке выравнивания j:

$$\text{ДИС}(j) = \frac{1}{N} \left(\sum_{i=1}^{k} \log x_i! - \log N! \right) \quad (3).$$

Число различных колонок с фиксированными частотами нуклеотидов как в колонке j при выравнивании N последовательностей равно:

$$I_N = \frac{N!}{x_1! .. x_k!} = e^{-N \cdot \text{ДИС}(j)} \quad (4).$$

Для больших x_i это число может быть приближено с использованием информационного содержания колонки I_s по Шнайдеру [Schneider и др., 1986]:

$$I_N = 2^{N(2-I_s)} \text{ где } I_s = 2 + \sum_{i=1}^{k} p_i \log_2 p_i \quad (5),$$

а $\{p_i\}, i\epsilon\{A, C, G, T\}$ – вероятности нуклеотидов в колонке.

ДИС не учитывает информацию о фоновых частотах нуклеотидов. В то же время, объединив формулы (2) и (3) мы получим:

$$\log P(x_1, .., x_k|Q, j) = -N \cdot \text{ДИС}(j) + \sum_{i=1}^{k} x_i \log q_i \quad (6).$$

Здесь первый член отражает «собственную негомогенность» колонки, а второй член описывает насколько наблюдаемое распределение нуклеотидов в колонке отличается от ожидаемого. Чтобы понять смысл формулы (6) следует рассмотреть аппроксимацию (2) для больших x_i. В этом случае можно использовать формулу Стирлинга:

$$\log n! \sim \frac{1}{2}\log(2\pi n) + n \log n - n \log e \quad (7).$$

Это позволяет развернуть (2) как:

$$\log P(x_1, .., x_k|Q, j) \sim -\frac{(k-1)}{2}\log 2\pi + \frac{1}{2}\log\frac{N}{x_1..x_k} + \sum_{i=1}^{k} x_i \log\frac{N}{x_i} + \sum_{i=1}^{k} x_i \log q_i.$$

Введем специфичную для колонки j оценку частот нуклеотидов $\tilde{p}_i = \frac{x_i}{N}$ и оставим только члены, пропорциональные N:

$$\log P(x_1, .., x_k|Q, j) \sim -N \sum_{i\epsilon\{A,C,G,T\}} \tilde{p}_i \log \tilde{p}_i + N \sum_{i\epsilon\{A,C,G,T\}} \tilde{p}_i \log q_i =$$

$$= -N \sum_{i\epsilon\{A,C,G,T\}} \tilde{p}_i \log\frac{\tilde{p}_i}{q_i} \quad (8).$$

Для больших x_i для конкретной колонки наблюдаемые нормализованные частоты $\tilde{p}_i$ сходятся к вероятностям нуклеотидов $\{p_i\}, i\epsilon\{A, C, G, T\}$ и формула (8) может быть переписана через расстояние Кульбака-Лейблера от Q до P [Kullback, Leibler, 1951]:

$$D_{KL}(P||Q) = \sum_{i\epsilon\{A,C,G,T\}} p_i \log\frac{p_i}{q_i} \quad (9).$$

Соответственно:

$$\log P(x_1, .., x_k|Q, j) \sim -N D_{KL}(P||Q) \quad (10).$$

Это позволяет переписать формулу (6):

$$\log P(x_1, .., x_k|Q, j) \sim -N \cdot \text{КДИС}(j) \quad (11),$$

где КДИС выполняет роль дискретного аналога расстояния Кульбака-Лейблера:

$$\text{КДИС}(j) = \text{ДИС}(j) - \sum_{i\epsilon\{A,C,G,T\}} \frac{x_i}{N} \log q_i \quad (12).$$

Интересно, что ДИС всегда отрицательно с максимумом в нуле. В то же время вычитаемый член также всегда отрицателен (поскольку $q_i \leq 1$). Следовательно, значение КДИС может быть как положительным, так и отрицательным.

В случае мотива связывания длины m (т.е. выравнивания «ширины» m), оценка качества мотива получается суммированием значений, полученных для всех колонок выранивания:

$$\text{КДИС} = \sum_{j=1}^{m} \text{КДИС}(j) \quad (13).$$

Именно это число ChIPMunk максимизирует при выборе оптимального мотива. Важно отметить, что отсчеты нуклеотидов $\{x_1, .., x_k\}$ могут быть вещественными числами (например, чтобы явно учитывать «достоверность» конкретного олигонуклеотида или последовательности), если вместо подсчета факториалов пользоваться аппроксимацией гамма-функции Эйлера. ChIPMunk использует формулу Стилтьеса [Abramowitz, Stegun, 1972] и это позволяет сохранить критерий оптимальности выравнивания при наличии априорных оценок достоверности (также называемых весами, не следует путать с весами – элементами ПВМ) или предположительных предпочитаемых позиций как для последовательностей обучающей выборки, так и для самого мотива.

Масштабирование информационного содержания

В оригинальном определении КДИС есть некоторые технические неудобства: величина информационного содержания зависит и от длины мотива (чем длиннее мотив – тем выше-лучше КДИС), так и от числа выравниваемых последовательностей N. Нормализация на длину мотива тривиальна (достаточно на нее поделить). Чтобы осмысленно отмасштабировать КДИС конкретной колонки мотива на N, необходимо оценить допустимый диапазон значений КДИС. Максимальное значение достижимо для колонки, полностью (для всех N слов) состоящей из вхождений наиболее «редкой» буквы α с наименьшей фоновой вероятностью q_α, $\alpha = \text{argmin}_{i\epsilon\{A,C,G,T\}}(q_i)$: $\text{КДИС}_{max} = -\log q_\alpha$.

Для простоты ChIPMunk (как и сопутствующие инструменты визуализации мотивов в виде лого-диаграмм на основе КДИС) всегда использует равномерное распределение $q_{i\epsilon\{A,C,G,T\}} = 0.25$. Потенциально, в экстремальных случаях мотив может существенно обогащен избегаемыми – с точки зрения фоновых частот – нуклеотидами. Это может вызвать появление КДИС-значений больше 1 после нормализации, но на практике встречается только в специально созданных тестовых примерах или при прямом переносе мотивов между геномами с существенно отличающимся GC-составом.

Минимальное значение КДИС можно оценить если вспомнить, что ДИС колонки аппроксимируется как $\sum_\alpha p_\alpha \log p_\alpha$ и, следовательно, КДИС колонки примерно равен $\sum_\alpha p_\alpha \log \frac{p_\alpha}{q_\alpha}$, где p_α – это нормализованная наблюдаемая частота нуклеотида α в заданной колонке мотива, а q_α – фоновая частота. Колонки наихудшего осмысленного мотива в точности повторяют фоновое распределение. Таким образом, минимальный КДИС колонки должен быть равен нулю. Теоретически, возможно появление отрицательных значений КДИС, например, если мотив по каким-то причинам избыточно обогащен наиболее частыми нуклеотидами фонового распределения, что не имеет смысла в реальных практических приложениях.

Таким образом, итоговая формула для подсчета КДИС имеет вид:

$$\text{КДИС} = -\frac{1}{m \log q_\alpha} \sum_{j=1}^{m} \frac{1}{N} \left(\sum_{i\epsilon\{A,C,G,T\}} \left(\log x_{i,j}! - x_{i,j} \log q_i \right) - \log N! \right) =$$

$$= \frac{1}{mN \log q_\alpha} \sum_{j=1}^{m} \left(\log N! + \sum_{i\epsilon\{A,C,G,T\}} \left(x_{i,j} \log q_i - \log x_{i,j}! \right) \right),$$

где m – длина мотива, q_α – частота наиболее редкого нуклеотида α (фиксирована как 0.25 для масштабирования КДИС, используемого в ChIPMunk), N – полное число слов в выравнивании, $i\epsilon\{A, C, G, T\}$ – нуклеотид, $x_{i,j}$ – ненормализованная частота (число отсчетов) конкретного нуклеотида i в j-й колонке выравнивания. Значения КДИС в практических приложениях попадают в диапазон [0,1], а значения порядка 0.5 соответствуют хорошо выраженным паттернам.

В практической реализации для подсчета КДИС используется натуральный логарифм, хотя выбор основания роли не играет.

4.1.2.2. Общая структура алгоритма

Нуклеотидный алфавит является достаточно ограниченным, а обучающая выборка биологических последовательностей содержит статистические отклонения различной природы, обусловленные как биологическим происхождением последовательностей, так и техническими артефактами конкретного эксперимента. Поиск характерного выраженного паттерна с максимальным информационным содержанием в зашумленной выборке можно условно представить как задачу оптимизации, например, аналогичную поиску минимума на сложном ландшафте, см. **Рисунок 16**. Один из простейших способов поиска оптимума – градиентный спуск – будет соответствовать скатыванию шарика по градиенту ландшафта.

Для ChIPMunk аналогом скатывания по градиенту является итеративная жадная оптимизация весовой матрицы, когда на каждом проходе из каждой последовательности выбирается слово (или слова) с наивысшей (для этой последовательности) оценкой на текущей весовой матрице, и матрица перестраивается из выбранных слов; а процесс повторяется до сходимости.

Скатывание шарика по сложному ландшафту осложнено локальными особенностями ландшафта, локальными оптимумами. Первая половина решения этой проблемы: использование множества шариков, сбрасываемых в различные точки ландшафта, что позволяет получить доступ к различным локальным оптимумам. Для ChIPMunk это соответствует использованию множества стартовых весовых матриц, случайных или получаемых на основе заданных консенсусов. Вторая половина решения: упрощение ландшафта. Для ChIPMunk это соответствует случайному сэмплированию – выделению случайных подвыборок «учебной» выборки последовательностей. Возврат к полной выборке для весовых матриц, оптимизированных на подвыборке, позволяет получить полное выравнивание всех последовательностей, а оптимизация на полной выборке сходится значительно быстрее, поскольку частично оптимизированная на подвыборке матрица в большинстве случаев уже похожа на целевой паттерн. Выбор итогового оптимального выравнивания и мотива производится по наилучшему дискретному информационному содержанию.

Задача идентификации мотива как построения множественного локального выравнивания последовательностей имеет смысл только в одном случае: если в

заметной доле последовательностей (в идеале – во всех последовательностях) действительно присутствуют вхождения характерного паттерна. Возможность быстрого «чернового» поиска мотива в случайных подвыборках согласуется с этой идеей.

Формализация алгоритма ChIPMunk в виде псевдокода

```
константа ЧИСЛО_ПОПЫТОК = 100 (по умолчанию)
константа ЧИСЛО_ШАГОВ = 10 (по умолчанию)
константа ЧИСЛО_ИТЕРАЦИЙ = 1 (по умолчанию)
константа МАКСИМАЛЬНЫЙ_КДИС_КОЛОНКИ = -log(0.25)
переменная наилучший_мотив

цикл1 ЧИСЛО_ПОПЫТОК раз

  переменная мотив = создать_случайную_весовую_матрицу
  переменная текущий_наилучший_мотив

  цикл2 ЧИСЛО_ШАГОВ раз

    переменная подвыборка = сэмплировать(все_последовательности)
    оптимизировать(мотив, подвыборка) ЧИСЛО_ИТЕРАЦИЙ раз
    оптимизировать(мотив, все_последовательности) до сходимости
    если КДИС(мотив) > КДИС(текущий_наилучший_мотив) то
       текущий_наилучший_мотив = мотив
    если КДИС(мотив) + МАКСИМАЛЬНЫЙ_КДИС_КОЛОНКИ < КДИС(наилучший_мотив) то
       мотив = наилучший_мотив

  цикл2 конец

  если КДИС(текущий_наилучший_мотив) > КДИС(наилучший_мотив) то
     наилучший_мотив = текущий_наилучший_мотив

цикл1 конец
```

МАКСИМАЛЬНЫЙ_КДИС_КОЛОНКИ используется здесь как порог при замене текущего мотива (во внутреннем цикле) на наилучший известный из прошлых итераций, чтобы избежать вычислительных затрат на оптимизацию неперспективных мотивов (слишком далеких от оптимума). Важно отметить, что циклы для различных стартовых мотивов могут выполняться параллельно в различных вычислительных потоках и обмениваться информацией о наилучшем решении. Текущая реализация ChIPMunk использует именно эту модель, которая на практике квази-линейно ускоряется с использованием 2-4 вычислительных потоков.

Узким местом алгоритма является оптимизация мотива – выбор наилучших вхождений из каждой последовательности (и для полного набора и для подвыборок) и перестройка весовой матрицы. Этот шаг потенциально может быть ускорен в

десятки раз при использовании подходящих структур данных для представления набора последовательностей или с помощью реализации на GPGPU.

Рисунок 16. Идея жадной оптимизации, используемой алгоритмом ChIPMunk. (*верхняя панель*) Различные локальные выравнивания соответствуют различным особенностям ландшафта. (*нижняя панель*) Сэмплирование выборки последовательностей аналогично упрощению ландшафта и позволяет выбраться из локальных оптимумов.

4.1.2.3. Оценка самосогласованности мотива для выбора порога отсечения

ChIPMunk ищет мотив, соответствующий безделеционному множественному локальному выравниванию с наивысшим дискретным информационным содержанием. В случае, когда все последовательности действительно содержат характерные вхождения мотива, этот подход позволяет выявить искомый паттерн. На практике в обучающей выборке всегда присутствует шум (т.е. последовательности, не содержащие сайтов связывания, соответствующих основному мотиву). Необходимо сегментировать выравнивание на «сигнал» и «шум». Для этого ChIPMunk строит серию весовых матриц из наилучших $1,2,\dots,n$ слов, последовательно выбираемых из ранжированного по оценкам ПВМ полного списка из N слов. Идея состоит в том, что «сигнал» расположен в верхней части списка слов, и, если n-е слово принадлежит «сигналу», его оценка по n-й весовой матрице должна быть значительно лучше, чем оценка по N-й матрице (построенной по всем N словам выравнивания всех N последовательностей и включающей шумовую компоненту). Формализацией этой идеи «самосогласованности» слов мотива и ПВМ является максимизация следующей суммы:

$$\max_{n=1..N} \sum_{z=1}^{n} \left(\text{оценка}(\text{ПВМ}_n, \text{слово}_z) - \text{оценка}(\text{ПВМ}_N, \text{слово}_z)\right)$$

где слово_z взято из z-й последовательности в отсортированном списке.

Поскольку N-я и n-я матрицы построены из разного числа слов, их оценки S несравнимы напрямую, ведь $S_{\alpha,j} = \log\left(\frac{x_{\alpha,j}+cq_\alpha}{(z+c)q_\alpha}\right)$, где j пробегает по колонкам мотива и $\alpha\epsilon\{A,C,G,T\}$ и, см. также раздел о позиционно-весовых матрицах в «Обзоре литературы». Чтобы сделать оценки сравнимыми, для каждой матрицы используются псевдоотсчеты c, пропорциональные числу слов $c = Kz$:

$$S_{\alpha,j} = \log\left(\frac{x_{\alpha,j}+Kzq_\alpha}{(z+Kz)q_\alpha}\right) = \log\left(\frac{x_{\alpha,j}+Kzq_\alpha}{(K+1)zq_\alpha}\right) = \log\left(\frac{\frac{x_{\alpha,j}}{z}+Kq_\alpha}{(K+1)q_\alpha}\right) = \log\left(\frac{p_{\alpha,j}+Kq_\alpha}{(K+1)q_\alpha}\right),$$

где $p_{\alpha,j}$ является нормализованной на число слов частотой нуклеотида α в j-й колонке. Это позволяет иметь сравнимую абсолютную шкалу для оценок вне зависимости от числа слов. Большие значения масштабного коэффициента K соответствуют большим псевдоотсчетам, т.е. меньшим штрафам оценки за нуклеотиды, не встречавшиеся в выравнивании при построении ПВМ и,

соответственно, более мягкой сегментации (включения большего числа слабых сайтов в выравнивание). Наиболее жесткая сегментация достижима для $K = 0$ (что обнуляет псевдоотсчеты). Практически удобным является значение $K = 1.0$.

4.1.2.4. Учет позиционных профилей

Последовательностям учебной выборки могут быть присвоены веса (априорные оценки достоверности, не следует путать с логарифмическими весами нуклеотидов в позиционно-весовой матрице). Эта операция не меняет алгоритма ChIPMunk, поскольку достоверности соответствующих слов выравнивания в этом случае напрямую используются при перестройке мотива по отсчетам нуклеотидов, которые становятся вещественными числами за счет домножения на значения весов. Для удобства веса (достоверности) всех последовательностей масштабируются, чтобы суммарная достоверность была равна числу последовательностей N. Опционально может применяться предварительная логарифмическая трансформация, чтобы исключить ситуацию, когда единственная последовательность с нереалистично большим значением заданной пользователем достоверности (например, экстремально высокий пик, полученный как артефакт ПЦР – в случае анализа ChIP-Seq данных) вносит драматически ошибочный вклад.

Аналогичным образом ChIPMunk учитывает и буквы в нотации IUPAC, соответствующие вариантам нуклеотидов в конкретной позиции последовательности; в этом случае дробные отсчеты назначаются конкретным нуклеотидам, например неизвестный нуклеотид **N** (не путать с числом последовательностей N) при построении вектора отсчетов для конкретной колонки мотива/выравнивания соответствует 0.25 **A**, 0.25 **C**, 0.25 **G** и 0.25 **T** (в качестве альтернативного варианта можно принимать дробные отсчеты равными частотам нуклеотидов в фоновом распределении; ChIPMunk для простоты ориентируется на равномерное распределение, как и при масштабировании КДИС).

Аналогичная идея взвешивания отсчетов $x_{\alpha,j}$ используется и для учета позиционных профилей последовательностей:

$$\alpha\epsilon\{\mathbf{A},\mathbf{C},\mathbf{G},\mathbf{T}\}: \quad x_{\alpha,j} = \sum_k w_k \cdot \text{профиль}_k[v_k + j] \cdot \delta\big(\alpha, \text{последовательность}_k[v_k + j]\big),$$

$$\alpha = \mathbf{N}: \quad x_{\mathbf{N},j} = \sum_k w_k \cdot \big(1 - \text{профиль}_k[j]\big),$$

$$\forall j: \quad \sum_{\alpha \epsilon \{\mathbf{A},\mathbf{C},\mathbf{G},\mathbf{T},\mathbf{N}\}} x_{\alpha,j} = \sum_k w_k,$$

где k – индекс последовательности во входном наборе, j – номер колонки в выравнивании, v_k – положение-сдвиг слова, соответствующего выравниванию, в k-й последовательности, профиль$_k$ представляет собой вектор позиционных предпочтений «вдоль» каждой последовательности, предварительно линейно отмасштабированный в диапазон [0,1] с помощью коэффициента w_k, а индикатор δ является символом Кронекера. Таким образом, позиционные профили явным образом штрафуют информационное содержание мотивов, локализованных в своих низких участках. Чтобы избежать шума, вызванного локальными особенностями профилей, ChIPMunk дополнительно фильтрует (сглаживает) входные профили с помощью простого фильтра «максимум по окну», где размер окна выбирается равным длине мотива (числу колонок выравнивания).

Отсутствие позиционного профиля эквивалентно «плоскому» профилю – вектору из единиц. Априорные достоверности последовательностей w_k в случае наличия позиционных профилей принимаются равными высоте профилей и, для удобства сравнения информационного содержания, нормируются так, чтобы их сумма $\sum_i w_i$ была равна полному числу последовательностей. В случае, когда известны только позиции вершин пиков, ChIPMunk самостоятельно может сгенерировать «треугольный» профиль с максимумом 1 в вершине и минимумом 0.5 на краях пика; на практике треугольные профили часто оказываются достаточными для правильной локализации мотива.

ChIPMunk использует две базовые операции: выбор наилучшего «слова» в последовательности и перестройку весовой матрицы из этих слов. Перестройка матрицы описана выше, а для подсчета оценок при выборе наилучшего слова-вхождения мотива в последовательностях с профилями используется следующая формула:

$$\text{оценка}(\text{ПВМ},\text{слово}) = \sum_{j=1}^{m} S_{\text{слово}[j],j} \cdot \text{профиль}[j].$$

Здесь достоверности последовательностей не учитываются, поскольку их значения не влияют на выбор наилучшего слова внутри каждой конкретной последовательности.

4.1.2.5. Учет формы мотива

Как было замечено еще на заре анализа мотивов [Papp, Chattoraj, Schneider, 1993], распределение информационного содержания по колонкам выравнивания сайтов связывания факторов транскрипции часто имеет периодичность, схожую с размерами витка спирали ДНК. Одно из возможных объяснений состоит в том, что наиболее консервативные нуклеотиды в выравнивании могут соответствовать прямым контактам с аминокислотами белка с ДНК через большую бороздку [Pabo, Sauer, 1984; Schneider, 2002]. Это свойство используется программой HeliCis [Larsson, Lindahl, Mostad, 2007], где задан априорный периодический спэйсер между двумя участками двухбоксового мотива.

Использование позиционных профилей для последовательности помогает выделить мотив, соответствующий ожидаемой локализации. Использование позиционных профилей для самого мотива, в свою очередь, потенциально позволяет точнее «фазировать» распределение информационного содержания по колонкам и избежать равномерного размазывания информационного содержания и получения слабовыраженного паттерна.

ChIPMunk может использовать два варианта «формы» мотива: однобоксовую и двухбоксовую. Форма однобоксового мотива описывается как $\cos^2(\pi n/T)$, где $T = 10.5$ и соответствует шагу спирали ДНК, а n является относительной координатой в списке колонок, т.е. $n = 0$ в центре мотива. Во внутреннем цикле оптимизации весовой матрицы оценки колонок умножаются на значения вектора «формы» и колонки ближе к центру не штрафуются, в то время как колонки на границе витка, т.е. с $n = 5{,}6, -5, -6$, дают гораздо меньший вклад в оценку. Аналогичным образом вводится двухбоксовый вариант формы мотива как $\sin^2(\pi n/T)$.

Форма мотива является вектором той же длины m что и мотив и содержит значения от нуля до единицы, где единица не дает штрафа к оценке конкретной позиции, а ноль соответствует позиции не участвующей в подсчете оценки (например, это может быть спэйсер-разделитель между боксами).

Формализация использования формы мотива выглядит так:

$$\text{оценка(ПВМ,слово)} = \sum_{j=1}^{m} S_{\text{слово}[j],j} \cdot \text{профиль}[j] \cdot \text{форма}[j].$$

Форма мотива влияет только на выбор наилучшего слова в каждой последовательности, но не меняет перестройку весовой матрицы по выравниванию лучших слов.

4.1.2.6. Выбор оптимальной длины мотива

ChIPMunk итеративно перебирает ширину выравнивания (длину мотива) в заданном пользователем диапазоне либо снизу вверх (от коротких длин к более протяженным) либо сверху вниз и останавливается когда находит наиболее длинный из протестированных *сильный* мотив.

Эмпирическое определение сильного мотива базируется на двух пороговых значениях информационного содержания: T и t. T соответствует КДИС колонки, в которой равномерно распределены только 3 из 4 возможных нуклеотидов, т.е. распределение частот имеет вид $\left[\frac{N}{3}, \frac{N}{3}, \frac{N}{3}, 0\right]$. Порог t соответствует КДИС колонки, в которой любые два одинаково представленных нуклеотида встречаются в два раза чаще, чем другие два, также одинаково представленные:$\left[\frac{2N}{6}, \frac{2N}{6}, \frac{N}{6}, \frac{N}{6}\right]$. На лого-диаграммах мотивов в работе эти пороги показаны пунктирными линиями (см. например **Рисунок 8**).

Мотив считается состоящим из нескольких доменов (боксов), если одна или более внутренних колонок выравнивания имеют КДИС менее t. По определению, для сильного мотива в целом и для всех его боксов существуют колонки с КДИС не хуже T, а все крайние колонки имеют КДИС не хуже t. Для хорошо выраженных мотивов и больших выборок последовательностей часто достаточно более простого критерия, в котором проверяется только КДИС крайних колонок самого мотива. Такой критерий используется в динуклеотидной версии алгоритма, описанной в отдельном разделе ниже.

4.1.3. Результаты базового тестирования

Алгоритм ChIPMunk во-многом базируется на эмпирических соображениях и адекватность результатов идентификации мотивов требует подтверждения. В ходе тестирования в рамках оригинальной статьи [Kulakovskiy и др., 2010] мы использовали три источника данных ChIP-Seq для факторов транскрипции NRSF/REST [Johnson и др., 2007], GABP [Valouev и др., 2008] и химерного белка

саркомы Юинга EWS-FLI1 [Guillon и др., 2009] (для него из набора связываемых последовательностей были предварительно исключены GGAA-повторы). Для простоты сравнения с другими программами длины мотивов были фиксированы как 21, 12 и 11 для REST, GABP и EWS-FLI1, соответственно, а в качестве набора пиков ChIP-Seq были взяты наилучшие 100 (REST, GABP) и наилучшие 500 (EWS-FLI1). В сравнении принимали участие три программы: MEME [Bailey и др., 2009], SeSiMCMC [Favorov и др., 2005] и HMS [Hu и др., 2010]. Последовательности центрировались на вершинах пиков и подвергались обрезке, оставляя от 10 до 100% длины. В качестве простой оценки качества подсчитывалась доля пиков, обрезанных до 10% длины, содержащих вхождения весовых матриц с оценками не хуже среднего плюс три стандартных отклонения (на основании распределения для всех возможных олигонуклеотидов фиксированной длины). Параллельно оценивалось время исполнения программ (Intel Core i7 920, openSUSE Linux 11). Результаты для EWS-FLI1 представлены на **Рисунке 17**. Все программы успешно идентифицируют мотив в коротких последовательностях (что согласуется с частой практикой использования обрезанных вокруг вершины пиков для идентификации мотивов). С ростом длины последовательности только ChIPMunk с учетом позиционных профилей (peak mode) способен корректно идентифицировать мотив. Время исполнения ChIPMunk является наименьшим в широком диапазоне длин последовательностей. Результаты для REST и GABP слабее зависят от длины последовательностей, в особенности для REST, что объяснимо чрезвычайно хорошо выраженным мотивом.

Мы также использовали наиболее сложные среди трех протестированных наборов данные для EWS-FLI1 для оценки стабильности сходимости ChIPMunk с настройками по умолчанию (100 попыток – стартовых случайных ПВМ, 10 шагов сэмплирования, 1 итерация оптимизации ПВМ на сэмплированных данных). **Рисунок 18** иллюстрирует сходство получаемых мотивов с помощью аналога ROC-кривых (по оси X отложены *P*-значения наилучших вхождений).

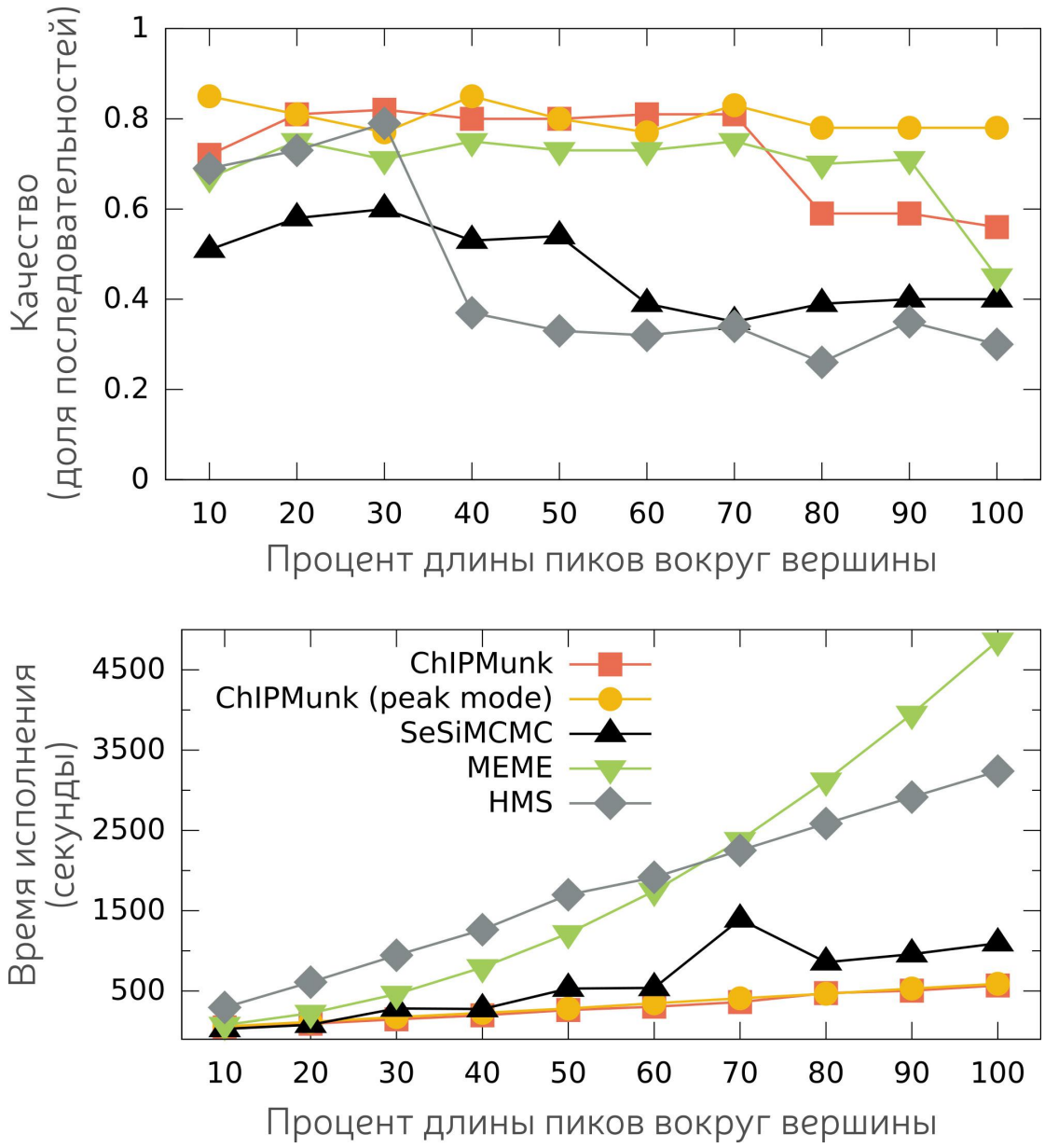


Рисунок 17. Эффективность идентификации мотивов в данных ChIP-Seq для белка EWS-FLI1 с помощью трех различных программ.

(*верхняя панель*) Время исполнения программ. (*нижняя панель*) Относительное качество мотивов, консенсусы соответствующих мотивов показаны в избранных точках, корректный консенсус имеет вид gacaGGAAatg. По оси X отложена относительная длина пиков, оставленная после обрезки вокруг вершин.

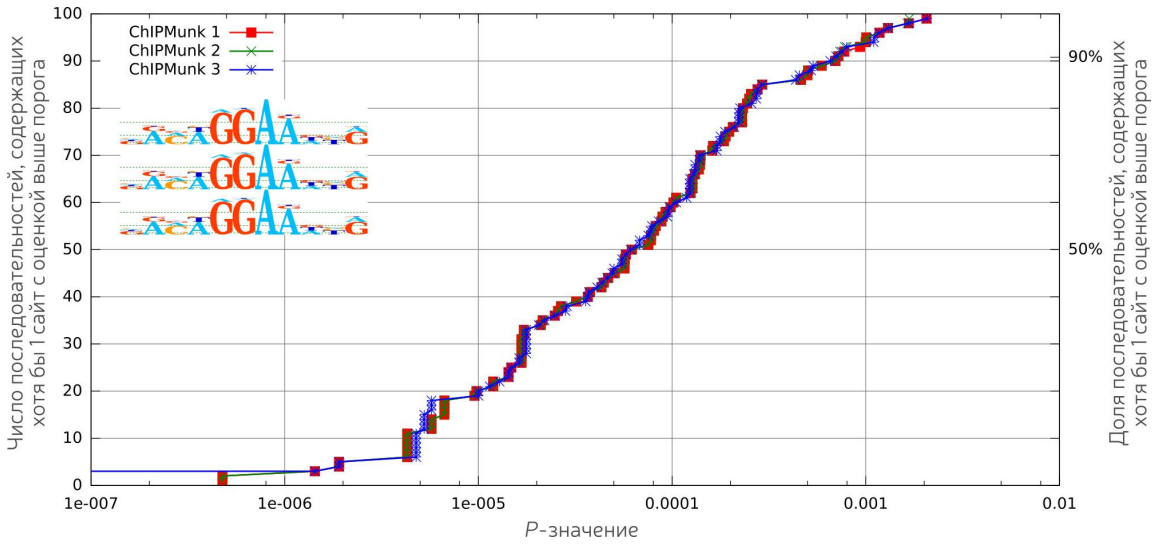


Рисунок 18. Упрощенный аналог ROC-кривых для трех независимых запусков ChIPMunk.

Шкала по оси X соответствует *P*-значениям наилучших вхождений. Значения по оси Y соответствуют доле последовательностей из тестового набора, для которых наилучшие вхождения имеют *P*-значения не выше порога (по оси X). Топ-100 пиков для EWS-FLI1 использованы для построения и тестирования мотивов. Лого-визуализации даны на врезке.

4.1.4. Практическое использование и ограничения применимости

Несмотря на ChIP-Seq-ориентированный дизайн метода, ChIPMunk успешно справляется с различными данными на основе секвенирования, например с результатами высокопроизводительных SELEX (что мы использовали при создании коллекции HOCOMOCO, см. ниже в разделе «Результаты»). ChIPMunk подходит для анализа сайтов связывания факторов транскрипции не только в эукариотических, но и в прокариотических геномах. Алгоритм успешно выступал в независимых тестированиях [Bi и др., 2011; Kuttippurathu и др., 2011; Ma и др., 2012], и был интегрирован в несколько программных комплексов, предназначенных для анализа мотивов и результатов высокопроизводительного секвенирования, в частности, BioUML[28], MotifLab[29] и Nebula[30]. Анализ ограниченного числа пиков (100-500) может быть проведен на персональном компьютере за небольшое время (минуты), анализ полноразмерных наборов данных (тысячи пиков длинной 1000 и более нуклеотидов) возможен, но требует уже более значительных вычислительных ресурсов. На практике чаще всего достаточно проанализировать только наиболее достоверную группу пиков.

ChIPMunk решает задачу построения наилучшего множественного локального выравнивания последовательностей. Это позволяет успешно идентифицировать основной (наиболее выраженный) мотив изучаемого ТФ, и, при итеративном использовании с расширением ChIPHorde (с поли**N**-маскированием вхождений или исключением последовательностей с надпороговыми вхождениями), найти еще 1-2 хорошо выраженных альтернативных варианта основного мотива или мотива кофакторов. Получаемые весовые матрицы хорошо отражают особенности сайтов связывания, а использование позиционной информации позволяет в большом числе случаев однозначно установить мотив целевого белка.

В то же время, ChIPMunk не является универсальным инструментом: он плохо справляется с идентификацией слабо-выраженных мотивов в шумных данных, и с выявлением полного спектра паттернов, перепредставленных в заданной выборке. Для решения этих задач лучше подходят альтернативные инструменты: от прямого

[28] BioUML Platform. http://biouml.org
[29] MotifLab is a general workbench for analysing regulatory sequence regions. http://motiflab.org
[30] Nebula web service. http://nebula.curie.fr/

поиска вхождений известных мотивов (motif finding), до идентификации паттернов *de novo* на основе словарного анализа, что успешно делает, например программа HOMER[31].

В заключение отметим, что ChIPMunk использует простые позиционно-весовые матрицы, построенные в предположении о независимости соседних нуклеотидов. Корреляции между соседними позициями в сайтах связывания можно учесть с помощью динуклеотидных позиционно-весовых матриц.

4.2. Построение расширенных моделей мотивов с учетом корреляций соседних позиций. Алгоритм diChIPMunk

Предположение о независимости позиций является одним из основных ограничений при использовании классических позиционно-весовых матриц в роли моделей сайтов связывания факторов транскрипции. Прямым расширением ПВМ является аналогичная весовая матрица большей размерности, учитывающая корреляции соседних нуклеотидов [Benos, Bulyk, Stormo, 2002; Gershenzon, Stormo, Ioshikhes, 2005]. Это выглядит осмысленным как с вычислительной точки зрения (большее число параметров позволяет точнее описать набор сайтов связывания), так и с точки зрения структурной организации контактов ДНК-белок и учета локальных особенностей спирали ДНК [Oshchepkov и др., 2004; SantaLucia, Hicks, 2004]. Исходная идея – о возможности модификации алгоритма ChIPMunk для работы с динуклеотидными матрицами – была выдвинута коллегами из Новосибирска: Дмитрием Ощепковым и Виктором Левицким. Ниже обсуждаются реализация и тестирование алгоритма diChIPMunk, опубликованного в работах [Kulakovskiy и др., 2013b; Kulakovskiy и др., 2013c].

Стоит отметить, что в последние годы появилось немало успешных примеров использования ChIP-Seq данных для построения и верификации еще более сложных моделей, явным образом учитывающих зависимости не только между соседними, но и между удаленными позициями [Keilwagen, Grau, 2015; Siebert, Söding, 2016]. Тем не менее, динуклеотидные ПВМ (диПВМ) diChIPMunk показывают достойные

[31] HOMER Motif Finding Tutorial - Homer Software and Data Download. http://homer.ucsd.edu/homer/motif/index.html

результаты и при сравнении с заметно более сложными моделями, что позволяет надеяться, что возможности их практического применения далеко не исчерпаны.

4.2.1. Переход к динуклеотидному алфавиту и построение динуклеотидных позиционно-весовых матриц

diChIPMunk полностью повторяет ChIPMunk в смысле базовых идей (общего дизайна алгоритма и процедуры итеративной перестройки мотива с сэмплированием обучающей выборки последовательностей), но все операции совершаются на динуклеотидном алфавите. В первую очередь, в динуклеотидный алфавит переводятся последовательности нуклеотидов учебной выборки. Каждый динуклеотид состоит из двух соседних нуклеотидов, и обратно, каждый нуклеотид, кроме первого и последнего в цепочке, принадлежит двум (перекрывающимся!) динуклеотидам. Динуклеотидный алфавит состоит из 16 «букв» {AA, AC, AG, .., TT}, но не любые пары динуклеотидных букв могут соседствовать в последовательности, записанной в динуклеотидном алфавите. Например, динуклеотидная последовательность AC-CG-GT соответствует нуклеотидной ACGT, но последовательность AC-GC-GT не имеет осмысленного аналога, который можно записать в нуклеотидах. То есть, в нашей постановке всевозможные последовательности на ACGT алфавите являются только подмножеством всех возможных динуклеотидных последовательностей. Эта заметная особенность не влияет на структуру алгоритма, поскольку в каждом конкретном случае мы всегда имеем дело с конкретным конечным подмножеством последовательностей, используемых для идентификации мотива. После конвертации последовательностей в динуклеотидный алфавит построение диПВМ и подсчет оценок слов происходят полностью аналогично однонуклеотидным матрицам: зависимости между соседними колонками выравнивания и весами в матрице реализуются неявно, через зависимости соседних букв последовательностей. Полученные таким образом динуклеотидные матрицы являются аналогом марковских моделей первого порядка.

Если ПВМ содержит всего «4 умножить на длину» весов-параметров (из которых «3 умножить на длину» по построению являются независимыми), то диПВМ – в 4 раза больше (веса 16 динуклеотидов в каждой колонке выравнивания). Это замедляет сходимость алгоритма и требует большего (по умолчанию – удвоенного) числа стартовых матриц и шагов сэмплирования для достижения достаточной

точности. На практике diChIPMunk оказывается медленнее ChIPMunk в 2-4 раза, что не мешает его использованию в массовом анализа ChIP-Seq данных.

4.2.2. Оптимальность выравнивания с учетом частот динуклеотидов и определение длины мотива

Для определения оптимальности выравнивания diChIPMunk пользуется дискретным инф. содержанием, переформулированным для динуклеотидного алфавита:

$$\text{КДИДИС} = \frac{1}{mN \log q_\alpha} \sum_{j=1}^{m} \left(\log N! + \sum_{i\epsilon\{AA,AC,AG,AT,..TT\}} \left(x_{i,j} \log q_i - \log x_{i,j}! \right) \right).$$

Обозначения повторяют таковые для ChIPMunk: m – длина мотива, q_α – частота наиболее редкого динуклеотида (фиксирована как $1/16 = 0.0625$ для простого масштабирования КДИДИС), N – полное число слов в выравнивании, $i\epsilon\{AA, AC, AG, AT, ..TT\}$ – динуклеотидная «буква», $x_{i,j}$ – ненормализованная частота (число отсчетов) конкретного динуклеотида i в j-й колонке выравнивания.

Отметим, что КДИДИС не учитывает зависимостей явным образом, и потенциально подсчет КДИДИС возможен для всех динуклеотидных последовательностей, в том числе, не имеющих записи в однонуклеотином алфавите. Тем не менее, это не мешает стратегии максимизации КДИДИС при выборе альтернативных выравниваний. Для реальных мотивов, построенных на реальных нуклеотидных последовательностях, соседние колонки будут зависимы в силу перекрытия составляющих их динуклеотидов.

При выборе длины мотива diChIPMunk пользуется упрощенным критерием сильного мотива, аналогично ChIPMunk: оценивается и сравнивается с порогом информационное содержание крайних колонок мотива. В качестве эмпирического порога взято КДИДИС колонки, в которой отсутствуют любые 2 динуклеотида, а оставшиеся 14 распределены равномерно.

ДИДИС используется аналогично ДИС для визуализации динуклеотидных мотивов в форме лого-диаграмм; в этом случае частоты нуклеотидов представлены в виде стандартного лого, а под ним приведены частоты динуклеотидов, захватывающих две соседние колонки. Пример визуализации для мотива связывания фактора транскрипции EBF1 (COE1) приведен на врезке.

4.2.3. Оценка результатов diChIPMunk с помощью операционных характеристик приемника

Для тестирования diChIPMunk мы использовали результаты ChIP-Seq проекта ENCODE для факторов транскрипции AP2A, HNF4A (с учетом позиционных профилей покрытия) и NRSF/REST (без априорной позиционной информации). Для упрощения процедуры, пики AP2A и HNF4A были обрезаны с двух сторон до достижения 10% от максимальной высоты профиля покрытия, а экстремально короткие пики (<25 пар нуклеотидов, п.н.) были исключены из рассмотрения. Всякий раз использовалась 1000 наилучших ранжированных по высоте пиков, пики четных и нечетных рангов использовались раздельно для идентификации и тестирования мотива, соответственно.

Для каждого фактора транскрипции рассматривались три модели, весовая матрица из TRANSFAC, весовая матрица ChIPMunk и динуклеотидная весовая матрица diChIPMunk. ChIPMunk и diChIPMunk строили мотив в диапазоне длин от 10 до 25 п.н. с использованием позиционного профиля для последовательностей (для AP2A и HNF4A) либо локального нуклеотидного или динуклеотидного состава (для REST). Для каждого фактора транскрипции строились ROC-кривые на основе наилучших вхождений мотивов в последовательности контрольной выборки (в качестве истинных положительных предсказаний), доля ложных положительных предсказаний оценивалась на основе *P*-значений (см. раздел о построении ROC-кривых ниже). Результаты представлены на **Рисунке 19**, значения AUC ROC приведены в **Таблице 1**.

Таблица 1. Значения площади под кривой (AUC) для ROC-кривых мотивов TRANSFAC, ChIPMunk (ПВМ) и diChIPMunk (диПВМ). Идеальный классификатор имеет AUC равной единице.

	AP2A	**REST**	**HNF4A**
диПВМ	0.936	0.975	0.962
ПВМ	0.915	0.957	0.957
TRANSFAC	0.874	0.940	0.782

Мотивы ChIPMunk во всех случаях оказались лучше TRANSFAC; причем прирост от перехода от моно- к динуклеотидным мотивам сопоставим с переходом от мотивов TRANSFAC к ChIPMunk. Интересное исключение – HNF4A – для которого разница

между моно- и динуклеотидными мотивами пренебрежимо мала, причем похожий эффект для HNF4A наблюдался и при сравнении других расширенных моделей с ПВМ [Levitsky и др., 2007].

Результаты ограниченного тестирования diChIPMunk хорошо воспроизводятся и для заметно более широкого спектра факторов транскрипции, что мы продемонстрировали при построении коллекции HOCOMOCO (см. «Результаты и обсуждение»).

4.2.4. Оценка качества динуклеотидных мотивов на основе локализации предсказанных сайтов связывания

В качестве альтернативного теста осмысленности диПВМ-мотивов мы изучили локализацию сайтов связывания в контрольных пиках (не использованных для построения мотива), предполагая, что более достоверные сайты должны быть идентифицированы ближе к вершинам.

В заметном числе пиков наилучшие вхождения ПВМ и диПВМ оказывались на удалении друг от друга более чем на 5 п.н. Для таких пиков мы рассмотрели локализацию наилучших предсказанных сайтов в различных окнах вокруг вершины (<25 п.н. до вершины, <50 п.н. до вершины, и т.д.; для REST вместо координат вершин использовались середины пиков). Результаты представлены на **Рисунке 20**: действительно, наилучшие с точки зрения диПВМ сайты точнее сфазированы с вершинами пиков. Интересно, что для HNF4A использование диПВМ не дает прироста, как и в тесте с ROC-кривой. Лого-диаграммы мотивов показаны на левой панели: для динуклеотидных мотивов под традиционной диаграммой дана аналогичная, демонстрирующая частоты динуклеотидов, захватывающих пары соседних колонок выравнивания.

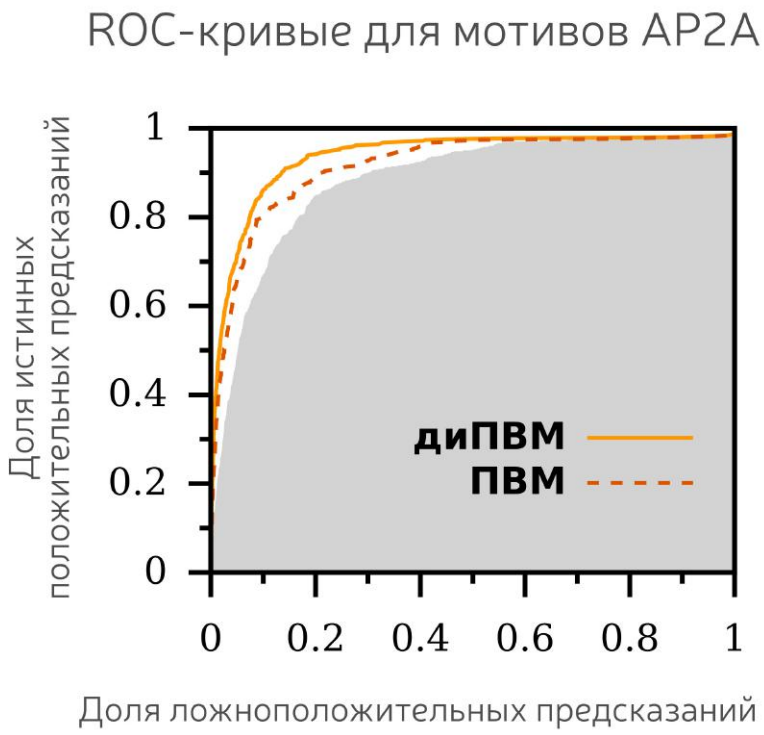


Рисунок 19. ROC-кривые для мотивов связывания фактора AP2A, полученные из TRANSFAC (ПВМ, выделена серая площадь под ROC-кривой), ChIPMunk (ПВМ, пунктирная линия), diChIPMunk (диПВМ, сплошная линия).

Для оценки доли истинных положительных предсказаний использован независимый поднабор пиков проекта ENCODE. Рисунок адаптирован из работы [Kulakovskiy и др., 2013b].

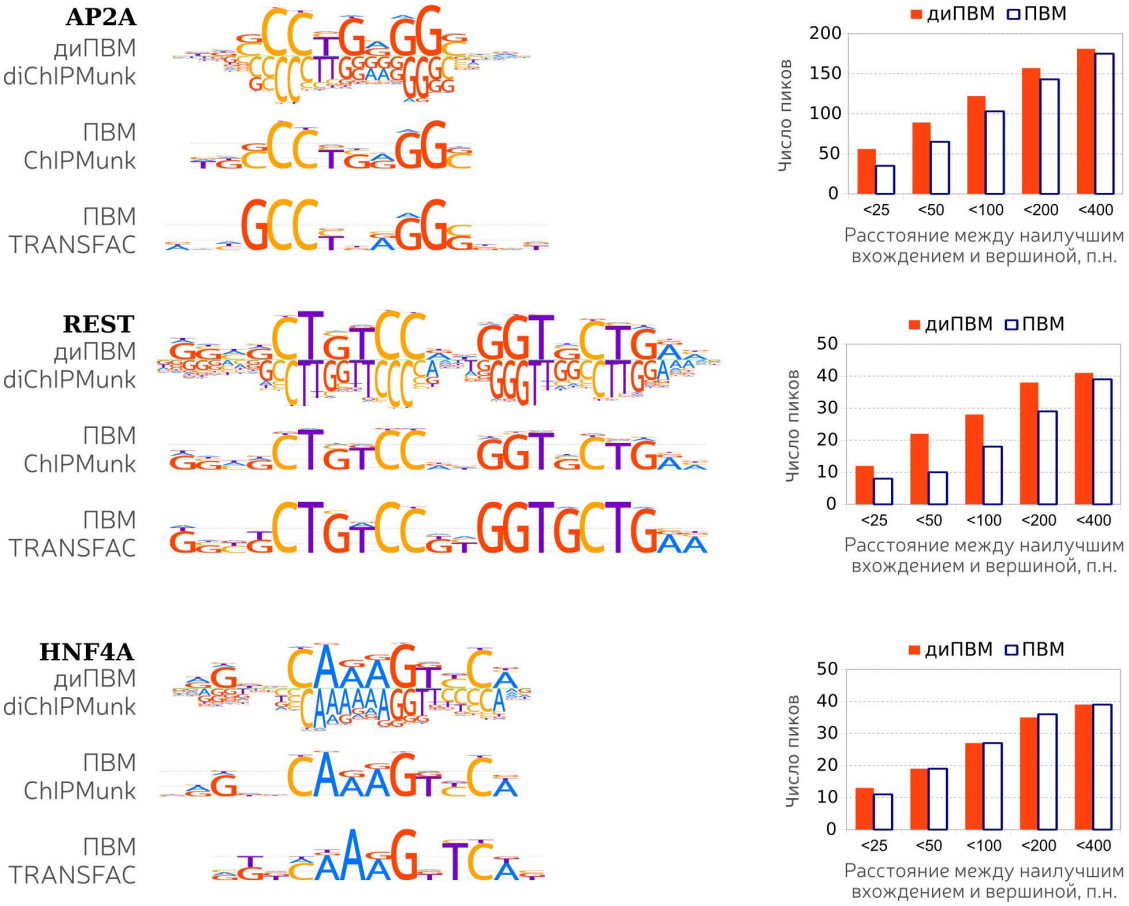


Рисунок 20. Фазирование мотивов AP2A относительно вершин пиков. (*левая панель*) Лого-визуализации мотивов ChIPMunk (ПВМ), diChIPMunk (диПВМ) и мотива из TRANSFAC (ПВМ). (*правая панель*) Кумулятивные распределения расстояний между наилучшими вхождениями ПВМ и диПВМ и вершинами пиков. Рассмотрены только пики контрольного набора с несовпадающими (на расстоянии не менее чем 5 п.н.) вхождениями ПВМ и диПВМ. Рисунок адаптирован из работы [Kulakovskiy и др., 2013b].

4.3. Естественная мера сходства мотивов

Позиционно-весовая матрица присваивает оценки всем словам-олигонуклеотидам конкретной длины. Выбор фиксированного порога для оценок позволяет ПВМ не только ранжировать слова, но и выделить подмножество наилучших, т.е. в случае анализа сайтов связывания факторов транскрипции условно разделить множество олигонуклеотидов на «сайты» и «не-сайты». Можно говорить о том, что ПВМ с пороговым значением оценки задает модель паттерна: ранжирование слов и определение «наилучшего» подмножества.

Последовательности, использованные при построении мотива, чаще всего являются результатом применения различных экспериментальных методов. Мотивы, идентифицированные в различных данных для одного и того же белка, часто похожи, но почти никогда не являются идентичными; это верно и для случая, когда один экспериментальный метод был использован для изучения нескольких членов одного структурного семейства. Таким образом, возникает задача сравнения мотивов не только с точки зрения точности классификаторов (построение ROC-кривых – как метод такого сравнения обсуждается ниже), но с точки зрения прямого сходства паттернов.

Стандартные методы сравнения ПВМ фактически сравнивают композицию отдельных колонок, т.е. напрямую используют элементы матриц и не учитывают ни значимостей целых слов ни пороговых значений оценок. Впервые «естественная» мера похожести матриц на основе целых слов была предложена в MoSta [Pape, Rahmann, Vingron, 2008] как асимптотическая ковариация вхождений мотивов в бесконечно длинную случайную последовательность. На наш взгляд, интуитивно более понятной является мера сходства, напрямую учитывающая число общих слов, т.е. имеющих надпороговые оценки для обоих мотивов.

Мы предложили использовать меру Жаккара для множеств слов, задаваемых мотивами как парами из ПВМ и фиксированных пороговых значений [Vorontsov, Kulakovskiy, Makeev, 2013]. На практике удобно, хотя и не обязательно, использовать пороги для сравниваемых ПВМ на уровне равных *P*-значений мотивов.

Мера Жаккара позволяет получить оптимальное выравнивание ПВМ, т.е. наилучший сдвиг одной ПВМ относительно другой, такой, что определяемые множества слов становятся наиболее похожими. Более того, сходство по Жаккару

позволяет задать метрическое пространство на однонуклеотидных моделях «ПВМ+порог» при условии, что слова равновероятны или порождаются случайной моделью Бернулли (т.е. как серии независимых испытаний – выбора буквы алфавита). Предложенная схема работает и для динуклеотидных весовых матриц, в свою очередь, при учете частот динуклеотидов и соответствующей динуклеотидной фоновой модели генерации слов (например, марковской модели 1-го порядка). Алгоритм и программа носят название MACRO-APE (MAtrix CompaRisOn by Approximate P-value Estimation, сравнение матриц с использованием приблизительной оценки *P*-значений), а формализация идей на примере мононуклеотидных весовых матриц представлена ниже.

4.3.1. Сходство мотивов по Жаккару

Пара из ПВМ и порога, которую мы здесь называем *моделью*, задает конечное подмножество среди всех олигонуклеотидов конкретной длины. Рассмотрим две модели, X и Y, задающие два набора слов X и Y одинаковой длины. Мера Жаккара между моделями будет иметь вид:

$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$, т.е. 

,

где |X| это размер множества – число слов в X, заданном моделью X (N на врезке соответствует числу элементов множества). То есть, $J(X,Y)$ задает долю слов, соответствующих обеим моделям (т.е. имеющих надпороговые оценки для обеих ПВМ) среди большего набора слов, распознаваемого любой из двух моделей.

Расстояние Танимото $D(X,Y) = 1 - J(X,Y)$ задает метрическое пространство для множеств слов фиксированной длины [Lipkus, 1999]. |X| и |Y| можно оценить с использованием существующих методов оценки *P*-значений мотивов при фиксированном пороге [Touzet, Varré, 2007]. Знаменатель можно переписать как $|X \cup Y| = |X| + |Y| - |X \cap Y|$, то есть остается подсчитать $|X \cap Y|$. Эта задача решается путем построения двумерного распределения пар оценок ПВМ, похожая техника была использована и в MoSta [Pape, Rahmann, Vingron, 2008].

Основным приложением мотивов в нашей работе являются сайты связывания факторов транскрипции, которые могут иметь произвольную ориентацию на двуцепочечной ДНК. Следовательно, оптимальное выравнивание ПВМ, помимо сдвигов, предполагает обратно-комплементарное преобразование. При сравнении

короткой и длинной ПВМ, «висящие» позиции для короткой матрицы могут быть заняты любыми нуклеотидами; для больших сдвигов матриц друг относительно друга висящие позиции у сравниваемых матриц будут появляться с противоположных сторон. Похожесть двух моделей мы определяем как наибольшую похожесть среди всех возможных выравниваний (сдвигов и обратно-комплементарного преобразования). Расширенное таким образом определение J не портит метрические свойства $D = 1 - J$ [Vorontsov, Kulakovskiy, Makeev, 2013].

4.3.2. Формализация позиционно-весовых матриц, P-значений мотивов и строгое определение меры сходства

Рассмотрим последовательность нуклеотидов (букв), записанную в алфавите $A = \{A, C, G, T\}$ и ПВМ M, матрицу 4 на m:

$M = [M(\alpha, i)]_{4 \cdot m}$, где позиции мотива соответствуют номерам колонок, символы алфавита ДНК – номерам строк, а m – длина мотива (например, длина сайта связывания; она же – ширина исходного выравнивания, использованного при построении весовой матрицы).

Элементы матрицы $M(\alpha, i) \epsilon \mathbb{R}$ представляют собой оценки букв $\alpha \epsilon A$ для всех позиций $1 \leq i \leq m$. Для каждого слова $\omega = \omega_1 .. \omega_m$ из A^m матрица M определяет оценку (скор):

$S(\omega, M) = \sum_{i=1}^{m} M(\omega_i, i)$.

Для фиксированного порога t, ПВМ определяет вхождение мотива в последовательность ζ в позиции n если $S(\zeta_n .. \zeta_{n+m-1}, M) \geq t$. Пара из ПВМ и порога t задает модель распознавания слов (например, сайтов связывания) и позволяет явным образом определить подмножество слов, соответствующих модели распознавания (т.е. разделить все множество m-меров на «сайты» и «не-сайты»):

$\Omega(M, t) = \{\omega \epsilon A^m : S(\omega, M) \geq t\}$.

P-значение ПВМ M для порога t – это вероятность $P(M, t)$ слова с оценкой не хуже t с точки зрения случайной модели генерации слов (в простейшем случае, как серии независимых испытаний на фиксированном распределении вероятностей букв-нуклеотидов – модель Бернулли):

$P(M, t) = \sum_{\omega \epsilon \Omega(M,t)} P(\omega)$,

где $P(\omega)$ это вероятность генерации конкретного слова ω.

Аналогично работе [Touzet, Varré, 2007] мы определяем распределение оценок $Q(M, s)$ как вероятность генерации слова с точной оценкой s:

$$Q(M, s) = \sum_{\omega: S(\omega, M) = s} P(\omega).$$

Если оценка s невозможна для ПВМ M, то $Q(M, s) = 0$. Знание распределения оценок позволяет определять P-значения:

$$P(M, t) = \sum_{s \geq t} Q(M, s).$$

4.3.2.1. Расширение и обратно-комплементарное преобразование ПВМ

Расширение ПВМ любым числом нулевых колонок слева или справа не меняет распределения оценок для любого порога оценки t.

С точки зрения сайтов связывания наилучшие вхождения мотивов могут располагаться на любой из двух цепей ДНК, соответственно, при поиске наилучшего выравнивания двух весовых матриц необходимо обратно-комплементарное преобразование:

$$\begin{matrix} \widetilde{M}(\mathrm{A}, i) = M(\mathrm{T}, m - i + 1); \widetilde{M}(\mathrm{T}, i) = M(\mathrm{A}, m - i + 1) \\ \widetilde{M}(\mathrm{C}, i) = M(\mathrm{G}, m - i + 1); \widetilde{M}(\mathrm{G}, i) = M(\mathrm{C}, m - i + 1) \end{matrix} \forall i$$

где $\widetilde{M}$ является обратно-комплементарным преобразованием M. При условии, что слова генерируются моделью Бернулли и вероятности нуклеотидов удовлетворяют второму правилу Чаргаффа, $p(\mathrm{A}) = p(\mathrm{T})$ и $p(\mathrm{C}) = p(\mathrm{G})$, обратно-комплементарное преобразование ПВМ не меняет распределения оценок [Vorontsov, Kulakovskiy, Makeev, 2013].

4.3.2.2. Выравнивание позиционно-весовых матриц

Рассмотрим две ПВМ, M_1 и M_2, потенциально представляющие мотивы разных длин m_1,m_2 и использованные для подсчета оценок в последовательности ζ в позициях j_1 и j_2, соотвественно. Будучи записанными с любым относительным сдвигом, эти две матрицы могут быть дополнены нулевыми колонками во всех невыровненных («висящих») позициях. Более формально, две ПВМ можно выровнять с заданным сдвигом, расширив M_1 нулевыми колонками в позициях, покрытых M_2, но не M_1, и обратно, расширив M_2 нулевыми колонками в позициях, покрытых M_1, но не M_2. Выровненные матрицы определяют мотивы одинаковой длины и задают оценки на одинаковом множестве слов фиксированной длины.

Для двух выровненных матриц M_1 и M_2 можно подсчитать P-значения для порогов t_1,t_2:

$$P(M_1, t_1) = \sum_{s \geq t_1} Q(M_1, s) = P\big(\Omega_1(M_1, t_1)\big),$$
$$P(M_2, t_2) = \sum_{s \geq t_2} Q(M_2, s) = P\big(\Omega_2(M_2, t_2)\big),$$

где Ω_1, Ω_2 – наборы слов, определяемые ПВМ M_1 и M_2 с соответстующими порогами t_1 и t_2.

Мера сходства наборов слов Ω_1, Ω_2 и соответствующих им моделей мотивов подчитывается как условная вероятность, что случайное слово ω имеет оценку для обеих матриц не хуже соответствующих порогов, при условии, что надпороговая оценка получена хотя бы для какой-то одной из двух ПВМ:

$$J1(\Omega_1,\Omega_2) = \frac{P(\{\omega\}:\omega \epsilon \Omega_1 \cap \Omega_2)}{P(\{\omega\}:\omega \epsilon \Omega_1 \cup \Omega_2)}.$$

В случае равномерного распределения вероятностей нуклеотидов $p(\alpha) = 0.25$ для всех $\alpha \epsilon A$, эта мера может быть упрощена до отношения числа слов, проходящих порог для обеих матриц, и числа слов, проходящих порог для любой из двух матриц:

$$J1(\Omega_1,\Omega_2) = \frac{|\Omega_1 \cap \Omega_2|}{|\Omega_1 \cup \Omega_2|},$$

что в точности соответствует сходству по Жаккару для двух наборов слов.

Расстояние $D1(\Omega_1,\Omega_2) = 1 - J1(\Omega_1,\Omega_2)$ задает метрическое пространство на взвешенных множествах слов, в нашем случае веса определяются как вероятности генерации слов фоновой моделью Бернулли (последовательностью независимых случайных испытаний на нуклеотидном алфавите с заданными частотами нуклеотидов).

Для выровненной пары ПВМ, определяющих множества слов Ω_1 и Ω_2, расширение ПВМ любым числом нулевых колонок не меняет значения $D1(\Omega_1,\Omega_2)$ [Vorontsov, Kulakovskiy, Makeev, 2013].

4.3.2.3. Итоговое определение меры сходства и расстояния между весовыми матрицами

Рассмотрим две невыровненные модели мотивов Ω_1 и Ω_2, представленные как ПВМ M_1 и M_2 возможно различных длин m_1, m_2 с порогами оценок t_1, t_2, соответствующих P-значениям $P_1 = P(M_1,t_1)$, $P_2 = P(M_2,t_2)$:

$\Omega_1 = \Omega(M_1,t_1)$ и $\Omega_2 = \Omega(M_2,t_2)$.

Каждая из двух ПВМ может определять слова на любой из цепей ДНК (т.е. к любой из двух ПВМ допустимо применить обратно-комплементарное преобразование), ПВМ могут быть выравнены с любым относительным сдвигом. Сходство ПВМ мы определяем как максимальное сходство, достижимое при рассмотрении всех возможных сдвигов и взаимных ориентаций:

$$J2(\Omega_1,\Omega_2) = \max_i \left(\max\left(J1_i(\Omega_1,\Omega_2), J1_i\left(\Omega_1,\widetilde{\Omega_2}\right)\right)\right);$$

и расстояние определяется аналогично:

$$D2(\Omega_1,\Omega_2) = \min_i \left(\min\left(D1_i(\Omega_1,\Omega_2), D1_i\left(\Omega_1,\widetilde{\Omega_2}\right)\right)\right).$$

Здесь $J1_i(\Omega_1,\Omega_2)$ задает сходство между моделями, определяемыми ПВМ M_1 и M_2, которые выровнены таким образом, 1-я колонка матрицы M_1 соответствует $(1+i)$-й колонке матрицы M_2, $1 - m_1 \le i \le m_2 - 1$, причем положительные значения i соответствуют расширению M_1 слева (и возможно M_2 расширенной справа), а $\widetilde{\Omega_2}$ представляет собой модель, полученную обратно-комплементарным преобразованием M_2. $J2$ определяет оптимальное выравнивание и взаимную ориентацию ПВМ M_1, M_2 для заданных порогов t_1, t_2.

Дистанция $D2(\Omega_1,\Omega_2) = 1 - J2(\Omega_1,\Omega_2)$ задает метрическое пространство на множестве моделей мотивов, определенных как пар из ПВМ и порогового значения оценки.

Подсчет размера или вероятности набора слов, распознанного одной или обеими моделями, требует построения распределения оценок, двумерного (для сравнения мотивов) или одномерного (для более простой задачи подсчета *P*-значений мотивов). Построение распределения не требует хранения или явного перебора всех слов, соответствующих мотиву, и проводится при помощи динамического программирования: используется словарь (хэш), подсчитывающий число m-меров, имеющих конкретное значение оценки (или пары оценок двух ПВМ); а длина слова m итеративно наращивается от единицы до длины мотива (или суммарной длины выровненных мотивов) с пересчетом ключей и значений словаря на каждом шаге. Для уменьшения размера словаря значения элементов ПВМ дискретизуются. Детали алгоритма изложены в работе [Vorontsov, Kulakovskiy, Makeev, 2013].

4.3.3. Практическое тестирование

Интересный практический вопрос – насколько схожи известные мотивы для одного и того же фактора транскрипции, представленные в существующих базах данных. В частности, эта информация позволяет оценить ожидаемый уровень сходства новых мотивов, идентифицированных в разных типах экспериментальных данных или разными программами.

Мы выбрали 85 пар ПВМ для факторов транскрипции, присутствующих одновременно в коллекциях JASPAR и HOCOMOCO. С помощью MACRO-APE было оценено сходство моделей для нескольких *P*-значений (для простоты всякий раз пороги двух ПВМ подбирались для одинакового *P*-значения; это не обязательно, но гарантирует сравнимый размер множества слов для каждого из тестируемых мотивов).

На **Рисунке 21** показано распределение значений сходства Жаккара для пар мотивов. Позитивный факт состоит в том, что модели для одного фактора транскрипции действительно более похожи, чем все прочие пары, причем в среднем сходство моделей слабо зависит от *P*-значения (т.е. от выбора порогов оценок). С другой стороны, абсолютный уровень сходства между парами моделей для конкретного фактора транскрипции довольно низок – лишь 30-50% слов из объединенного множества распознаются обоими мотивами.

На рисунке **Рисунке 22** показано среднее и стандартное отклонение для сходства моделей одного ТФ, подсчитанных для разных *P*-значений. Интересно, что дисперсия сходства слабо отличается для средних и высоких *P*-значений. В то же время на практике часто используются низкие *P*-значения, чтобы минимизировать ложноположительные предсказания ПВМ. В этой области дисперсия сходства выше, а среднее сходство падает, конкретное значение существенно зависит от фактора транскрипции (см. **Рисунок 23**). Для ряда факторов транскрипции это неожиданно означает, что наилучшие предсказания ПВМ наихудшим образом согласуются между разными моделями. Интересно, что для белка CTCF лого-диаграммы практически неотличимы, а сходство по Жаккару достигает лишь 0.6, что соответствует 60% общих предсказаний ПВМ среди всех слов, распознанных любой из моделей, или порядка 20% уникальных предсказаний для каждой из моделей.

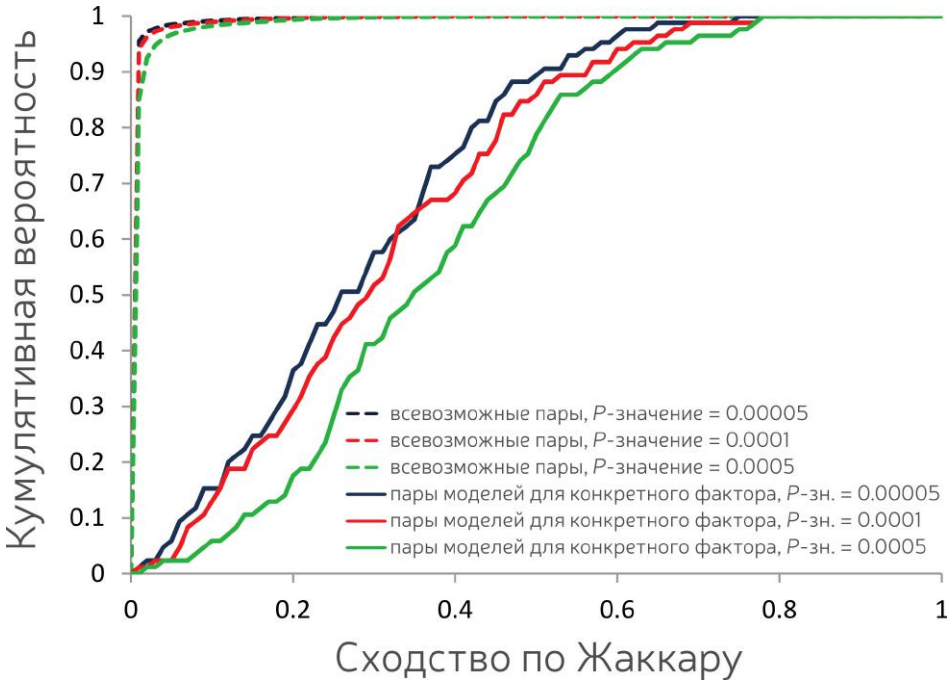


Рисунок 21. Кумулятивное распределение сходства моделей сайтов связывания для факторов транскрипции, представленных одновременно в HOCOMOCO и JASPAR. Всевозможные пары из 170 моделей использованы как нейтральный контроль (пунктирные линии). Различные цвета кривых соответствуют различным *P*-значениям. Рисунок адаптирован из работы [Vorontsov, Kulakovskiy, Makeev, 2013].

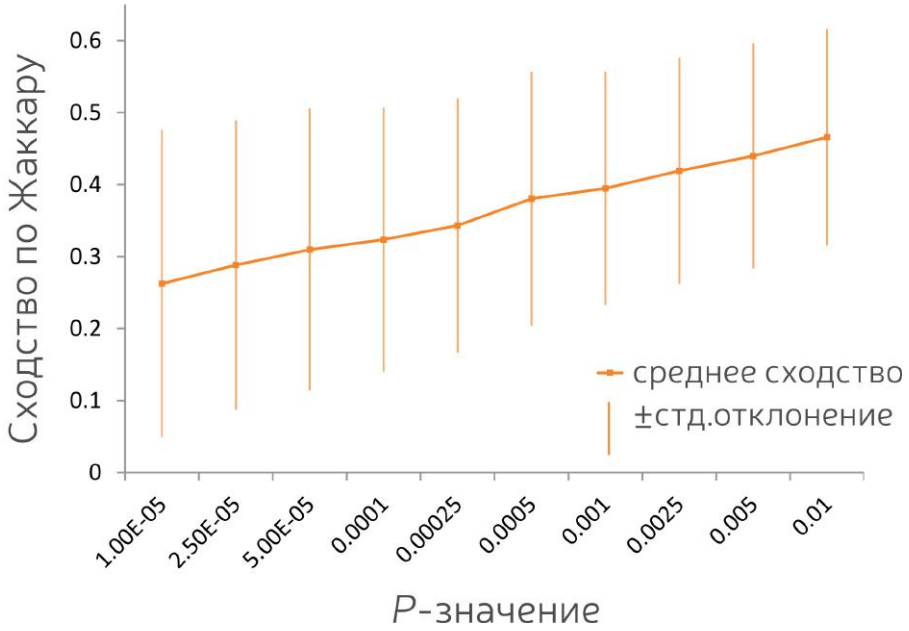


Рисунок 22. Среднее и стандартное отклонение значений сходства по Жаккару между моделями сайтов связывания для факторов транскрипции, представленных одновременно в HOCOMOCO и JASPAR (85 пар моделей).

Рисунок адаптирован из работы [Vorontsov, Kulakovskiy, Makeev, 2013].

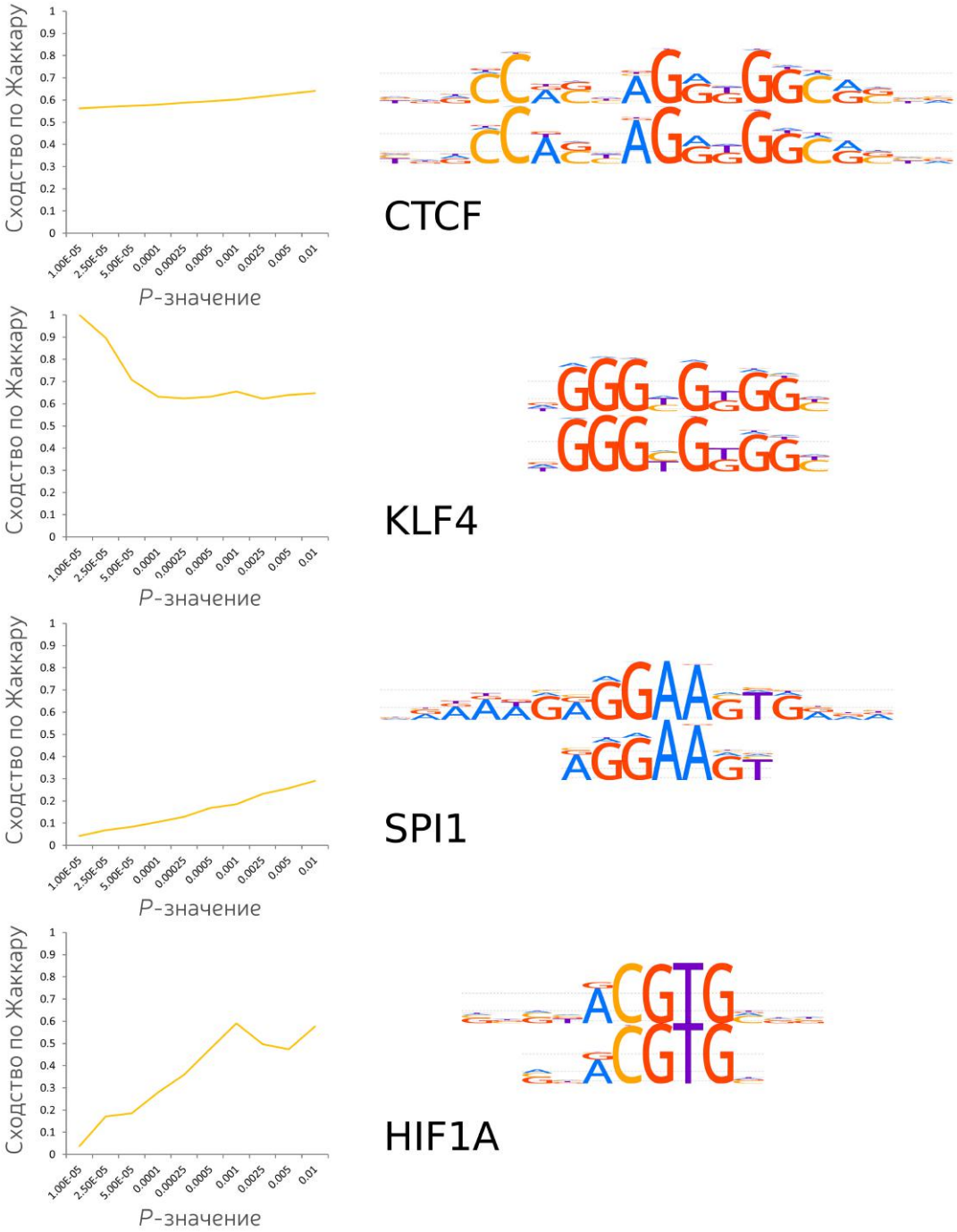


Рисунок 23. (*левые панели*) Сходство по Жаккару и (*правые панели*) лого-визуализации для моделей сайтов избранных факторов транскрипции, представленных в HOCOMOCO и JASPAR.

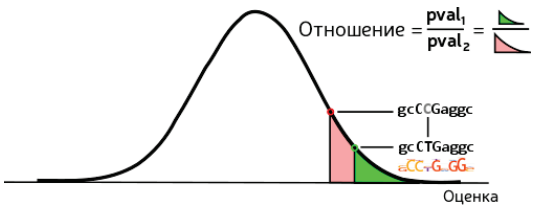
Ось X соответствует различным *P*-значениям. Рисунок адаптирован из работы [Vorontsov, Kulakovskiy, Makeev, 2013].

4.4. Сопутствующие методы анализа мотивов

4.4.1. Аннотация регуляторных вариантов в сайтах связывания факторов транскрипции. Алгоритм и программа PERFECTOS-APE

Геномные варианты, локализованные в некодирующих областях, потенциально могут оказывать влияние на экспрессию генов. Одним из возможных механизмов может быть изменение аффинности сайтов связывания факторов транскрипции и, как следствие, изменение активности соответствующего промотора или энхансера. Предсказать или объяснить такие эффекты может как общегеномная аннотация (например, карта доступности хроматина), так и локальный анализ мотивов, вхождения которых перекрываются с геномным вариантом.

Идея нашего метода PERFECTOS-APE (Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation), в сущности, повторяет существующие работы [Macintyre и др., 2010; Manke, Heinig, Vingron, 2010], но реализация позволяет использовать как моно- так и динуклеотидные весовые матрицы в качестве моделей мотивов. Алгоритм оценивает *P*-значения предсказанных сайтов связывания в зависимости от аллельных вариантов конкретной геномной позиции. Отношение *P*-значений (см. врезку) и минимальное *P*-значение служат для выбора наиболее интересных наблюдений. Интересно, что связь между *P*-значением мотива и энергией специфического связывания (discriminative energy) обсуждалась еще в приложении к классической работе [Berg, 1987], откуда следует, что для однонуклеотидной замены в мотиве связывания логарифм отношения *P*-значений, соответствующих альтернативным вариантам, близок к разности соответствующих энергий. Пример применения PERFECTOS-APE к анализу реальных полиморфизмов представлен на **Рисунке 24**.



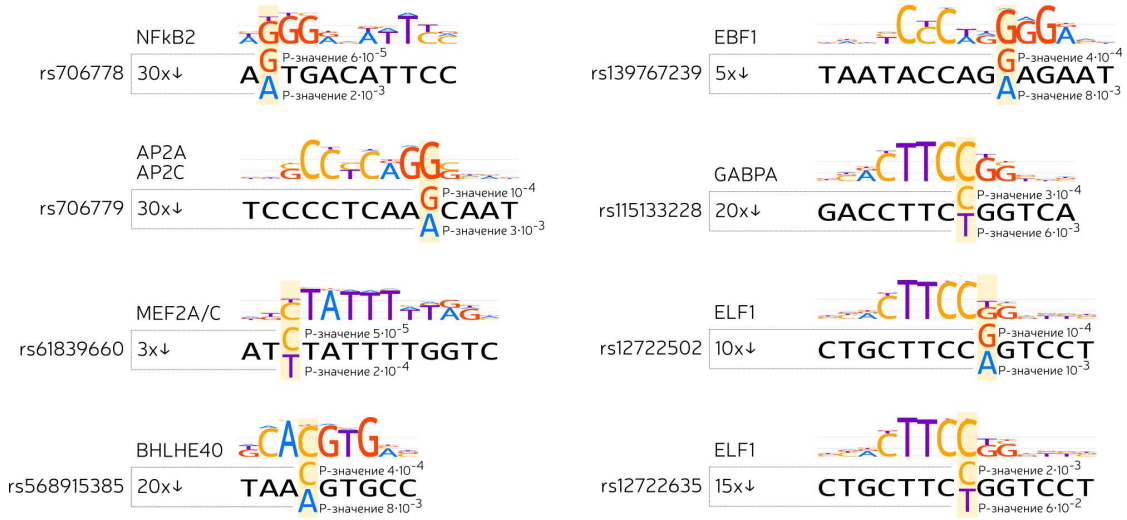


Рисунок 24. Полиморфизмы первого интрона гена IL2RA, ассоциированные с аутоиммунными заболеваниями, и соседние генетически сцепленные однонуклеотидные варианты изменяют аффинность сайтов связывания факторов транскрипции.

Идентификаторы полиморфизмов приведены по dbSNP, показаны *P*-значения для альтернативных аллелей, отношения *P*-значений и лого-визуализации мотивов. Рисунок адаптирован из работы [Schwartz и др., 2017].

4.4.2. Поиск вхождений мотивов в нуклеотидных последовательностях. Алгоритм и программа SPRy-SARUS

Практические приложения мотивов требуют поиска вхождений в заданных последовательностях. Эта достаточно стандартная задача, которая для мононуклеотидных матриц решается существующими инструментами. Для расширенных моделей (учитывающих зависимости между нуклеотидами) быстрые методы поиска вхождений появились только в последние годы [Korhonen и др., 2016]. Для моно- и динуклеотидных весовых матриц мы разработали свой инструмент SPRy-SARUS (Straightforward yet Powerful Rapid SuperAlphabet Representation Utilized for motif Search, супералфавитное представление для поиска мотивов). Этот простой метод использует супералфавитный подход, предложенный ранее в работах [Korhonen и др., 2009]. Супералфавитный метод позволяет уменьшить число операций при наивном последовательном сканировании последовательности путем замены алфавита: переход от моно- к динуклеотидам, или от ди- к тринуклеотидам. Полезно понимать, что это отличается от подхода, использованного в diChIPMunk для диПВМ: соседние буквы в супералфавитной записи не перекрываются, а выигрыш по времени достигается за счет манипуляции с представлением матрицы (которая для супералфавита содержит оценки не для отдельных букв, а сразу суммарные для пар букв, и, фактически, вдвое уменьшает эффективную длину мотива и необходимое число операций при подсчете оценок).

4.4.3. Сравнение качества распознавания сайтов связывания с помощью ROC-кривой. Статистическая оценка ожидаемой доли ложноположительных предсказаний

ROC-кривая показывает зависимость доли истинных положительных предсказаний от доли ложноположительных предсказаний. Подсчет числа истинных положительных предсказаний возможен на основе независимой выборки экспериментально определенных сайтов связывания, либо путем кросс-валидации данных различных экспериментов, либо путем деления одной выборки на поднаборы для обучения и тестирования. В нашей работе мы повсеместно используем схему, когда пики ChIP-Seq ранжируются по высоте или значимости, и пики четных/нечетных рангов раздельно используются для идентификации/тестирования мотива. В качестве истинных положительных предсказаний мы подсчитываем пики, для

которых наилучшее вхождение мотива имеет оценку не хуже пороговой (порог оценки выполняет функцию свободного параметра, т.е. перебор различных порогов оценки соответствует движению вдоль ROC-кривой).

Более сложный вопрос – подсчет ложноположительных предсказаний. За редким исключением, экспериментальная информация об олигонуклеотидах, достоверно не связываемых белком, оказывается недоступной. Типичная вычислительная схема: случайное перемешивание нуклеотидов в последовательностях контрольной выборки (т.е. содержащих сайты связывания). Альтернативный вариант: сэмплирование близлежайших геномных районов. Мы предлагаем третий подход, который позволяет получить стабильный аналог доли ложноположительный предсказаний.

Для каждого порогового значения оценки мотива (в виде ПВМ или диПВМ) мы подсчитываем P_s как вероятность случайной встречи хотя бы 1 надпорогового вхождения ПВМ в случайную последовательность ДНК фиксированной длины L, например, выбранной как медиана длин пиков ChIP-Seq для конкретного белка (что обычно соответствует значениям около 300-500 п.н.). P_s подсчитывается на основе P-значения мотива как вероятность получения ПВМ-оценки не хуже порога для случайного слова хотя бы в одной позиции случайной двуцепочечной последовательности ДНК, в предположении, что вхождения ПВМ (включая перекрывающиеся вхождения) являются независимыми, и их полное число удовлетворяет сложному (составному) распределению Пуассона:

$$P_s = 1 - (1 - P)^{2(L-l+1)},$$

где l соответствует длине мотива, а P-значение мотива (P), может подсчитываться с помощью MACRO-APE с учетом нуклеотидного или динуклеотидного состава (этот вариант использован при тестировании мотивов в последней версии коллекции HOCOMOCO). Для равновероятных нуклеотидов P_s соответствует доле всевозможных последовательностей длины L, в которых находится одно и более надпороговое вхождение мотива. Подсчет доли ложноположительных предсказаний с помощью P_s не является абсолютно точным, в силу предположения о независимости соседних вхождений, фиксированной длины последовательностей и того факта, что всевозможные последовательности включают в себя и истинно

связываемые белком, т.е. содержащие надпороговые вхождения мотива. Тем не менее, можно считать, что их число пренебрежимо мало в общем пуле всевозможных последовательностей, особенно для высоких порогов оценок (т.е. для малых значений P_S – в левой части ROC-кривой). Независимая проблема – трудность сравнения мотивов характерно разных длин, т.к. короткие мотивы имеют ограниченный диапазон собственных *P*-значений. Тем не менее, на практике построенные по предложенной схеме ROC-кривые являются достаточно мощным и удобным инструментом для сравнения мотивов как классификаторов.

4.5. Техническая реализация и доступность методов

Воспроизводимость является одним из проблемных моментов современных исследований. Для биоинформатики воспроизводимость реализуется через открытую реализацию и публикацию в сети Интернет основных программных средств, используемых в ходе анализа.

Все основные компьютерные методы, использованные в ходе этой диссертационной работы, реализованы в виде java-пакетов с открытым исходным кодом. Программная реализация ChIPMunk и diChIPMunk выполнена лично автором диссертации, первичная реализация SPRy-SARUS выполнена Анастасией Соболевой (Денисенко) под руководством автора диссертации в рамках дипломной работы, реализация MACRO-APE и PERFECTOS-APE выполнена Ильей Воронцовым в рамках совместной работы. Аннотации всех программ и ссылки на сайты представлены на портале autosome.ru[32]. Совместно с Ильей Воронцовым для основных программ разработан единый веб-интерфейс под общим названием «Оперный театр»[33]. Программы снабжены подробными инструкциями на английском языке и поставляются с открытым исходным кодом по лицензии WTFPL[34] ☺.

[32] autosome.ru Bioinformatics Tools. http://autosome.ru
[33] autosome.ru Opera House. http://opera.autosome.ru
[34] WTFPL. http://wtfpl.net

5. Результаты и обсуждение

Раздел посвящен описанию основных результатов диссертации (за исключением методических достижений, описанных в разделе «Материалы и методы»). Раздел включает описание авторской коллекции HOCOMOCO, систематического каталога мотивов связывания факторов транскрипции человека и мыши, и результаты использования коллекции и сопутствующих биоинформатических инструментов для анализа мотивов в различных задачах регуляторной геномики высших эукариот.

5.1. Коллекция HOCOMOCO: мотивы сайтов связывания факторов транскрипции человека и мыши

5.1.1. Построение базовой коллекции мотивов путем интеграции данных различных источников

Существует множество репозиториев-коллекций известных мотивов сайтов ДНК-связывающих белков, которые можно использовать при анализе последовательностей промоторов и энхансеров. Проблема практического использования коллекций состоит в «несистематической избыточности». Для конкретного фактора транскрипции даже в одной коллекции часто присутствует несколько похожих мотивов («профилей связывания»), построенных по различным экспериментальным данным и/или различными вычислительными методами[35]. В частности, страдает от присутствия избыточных моделей популярная ранее коммерческая коллекция TRANSFAC [Wingender и др., 2000]. Вполне разумна идея о том, что мотив связывания фактора транскрипции зависит от экспериментальных данных и биологических условий. Но этот тезис не может служить оправданием для использования недостоверных моделей, демонстрирующих не биологические свойства фактора транскрипции, а особенности конкретных экспериментов или методов идентификации мотивов. Трудно понять насколько корректно мотив описывает предпочтения связывания фактора транскрипции без прямой экспериментальной валидации или совместного вычислительного анализа различных источников данных.

[35] См. например разнообразие мотивов связывания фактора FoxA2 (информация актуальна на февраль 2017) в базе данных Transcription Factor Encyclopedia [Yusuf и др., 2012]. http://cisreg.cmmt.ubc.ca/cgi-bin/tfe/articles.pl?tfid=166&tab=tfbs#TF0000352

В этом разделе диссертации обсуждается построение коллекции HOCOMOCO (в первом публичном релизе – HOmo sapiens COmprehensive MOtif Collection, в последующем релизе – HOCOMOCO COmprehensive MOtif COllection)[36], содержащей мотивы связывания для сотен факторов транскрипции человека и мыши. В целях согласования терминологии, в HOCOMOCO мотивы называются моделями сайтов связывания, понимая под моделью паттерн плюс подсчитанные для различных фиксированных *P*-значений пороговые оценки.

Мотивы в HOCOMOCO были построены единообразным способом с помощью авторских программ, описанных в «Материалах и методах» и совместного анализа данных различных экспериментов по ДНК-белковому узнаванию. По построению, в коллекции отсутствует избыточность: для одного фактора транскрипции альтернативные мотивы допускаются только для случаев, когда различные режимы связывания подтверждены экспериментально (например, связыванию в формах мономера и димера может соответствовать пара из одно- и двухбоксового мотива).

5.1.1.1. Общие соображения о построении коллекции и идентификации мотивов

Одной из стартовых целей при создании коллекции было систематическое построение устойчивых «базовых» моделей для факторов транскрипции человека: мы были готовы пожертвовать детальным отражением особенностей конкретных экспериментальных условий или клеточных типов, чтобы для конкретного фактора транскрипции получить надежное описание основных предпочтений по узнаванию ДНК.

При наличии ограниченного набора разнородных экспериментальных данных процедура выявления мотивов может быть несогласованной и порождать похожие, но не идентичные паттерны. Полноценная вычислительная валидация (например, сравнительная оценка качества похожих мотивов с помощью ROC-кривых) возможна только при наличии достаточного объема независимых экспериментальных данных, не использованных при построении мотива (это было сделано на второй итерации создания коллекции, описываемой в следующем разделе).

---

[36] HOCOMOCO COmprehensive MOtif COllection. http://autosome.ru/HOCOMOCOS/

В общем случае, улучшить устойчивость процедуры выявления мотивов можно ограничив степени свободы при идентификации паттерна. При построении HOCOMOCO мы использовали возможности программы ChIPMunk: (1) прямую интеграцию данных различных экспериментальных методов [Kulakovskiy, Makeev, 2009], т.е. выявление мотива одновременно в полном наборе имеющихся разнородных нуклеотидных последовательностей с учетом их достоверности; (2) априорную информацию о положении сайта связывания (основываясь на высотах и формах пиков ChIP-Seq и существующей экспертной разметке регионов связывания, определенных догеномными методами); (3) априорную информацию о «форме» мотива (вписывая одно-/двухбоксовые мотивы в один/два витка спирали ДНК, см. раздел о ChIPMunk в «Материалах и методах»). Результаты вычислительной идентификации мотивов затем проходили стадию экспертного курирования.

Для удобства практического использования данные по ДНК-белковому взаимодействию (исходные последовательности сайтов или геномные районы) и результирующие мотивы однозначно связывали с идентификаторами белков по базе UniProt [The UniProt Consortium, 2012]. В целях увеличения объема обучающих выборок последовательностей, что особенно критично для малоизученных факторов транскрипции или при отсутствии данных высокопроизводительных экспериментов, мы объединяли информацию обо всех сайтах связывания для гомологичных факторов транскрипции позвоночных (по данным JASPAR и TRANSFAC).

5.1.1.2. Обзор источников данных

В первой версии коллекции мы сделали акцент на интеграции данных из различных источников, включая систематизированные базы данных:

(1) База данных JASPAR (доступная в 2012 году коллекция CORE VERTEBRATE). В анализ были взяты последовательности, использованные для построения моделей сайтов связывания; а готовые модели JASPAR использовались в сравнительном тестировании на ограниченном наборе ChIP-Seq данных (см. ниже);

(2) База данных TRANSFAC (релиз 2011.2, подраздел SITE, данные обработаны под рук. Владимира Байича[37] в KAUST, Саудовская Аравия, в рамках партнерского проекта). Были извлечены данные о регионах связывания, аннотированные в TRANSFAC по литературным данным; а готовые модели мотивов использовались в ходе сравнительного тестирования.

Для сайтов связывания, аннотированных в базах данных JASPAR и TRANSFAC, мы генерировали позиционный профиль, примерно указывающий проаннотированное положение сайта связывания в исходных фрагментах ДНК.

(3) Опубликованные к моменту создания коллекции предварительные результаты ChIP-Seq проекта ENCODE. Что интересно, наша работа была сделана и успешно опубликована до анализа мотивов, проведенного в рамках самого консорциума ENCODE [Kheradpour, Kellis, 2014]). Для ChIP-Seq мы восстановили позиционные профили или использовали значимости пиков как веса последовательностей.

(4) Первые массовые результаты HT-SELEX [Jolma и др., 2010]. Здесь мы использовали число прочтений олигонуклеотида как вес последовательности, что позволило сгруппировать идентичные прочтения и уменьшить объем выборки без потери информации.

Для некоторых факторов транскрипции были дополнительно проанализированы литературные источники (например, были добавлены результаты ChIP-chip эксперимента для белка p53 [Smeenk и др., 2008]).

Интересно, что TRANSFAC и ENCODE ChIP-Seq представляли наиболее несхожие источники данных: из TRANSFAC были извлечены ограниченные наборы сайтов связывания, определенные догеномными методами для широкого спектра факторов транскрипции, а из ChIP-Seq – сотни и тысячи сайтов связывания лишь для нескольких десятков белков.

[37] V. Bajic, Computational Bioscience Research Center
https://www.kaust.edu.sa/en/study/faculty/vladimir-bajic

В **Таблице 2** приведен краткий обзор использованных источников данных. В дальнейшем анализе участвовало 474 фактора транскрипции, для которых удалось найти более 4 идентифицированных последовательностей сайтов или регионов связывания. Диаграмма Венна, иллюстрирующая покрытие факторов транскрипции исходными данными, приведена на **Рисунке 25**.

Таблица 2. Обзор источников данных базовой версии HOCOMOCO.

Источник данных (число факторов транскрипции)	**Полное число послед-й**	**Медиана длин послед-й (п.н.)**	**Средняя длина послед-й (п.н.)**
TRANSFAC (442)	23199	24	26
JASPAR (108)	24692	204	207
ENCODE ChIP-Seq (Yale, 50)	96381	454	656
ENCODE ChIP-Seq (HudsonAlpha, 28)	65081	559	561
HT-SELEX (19) [Jolma и др., 2010]	19535	16	16
Прочие источники (29)	2655	1000	687

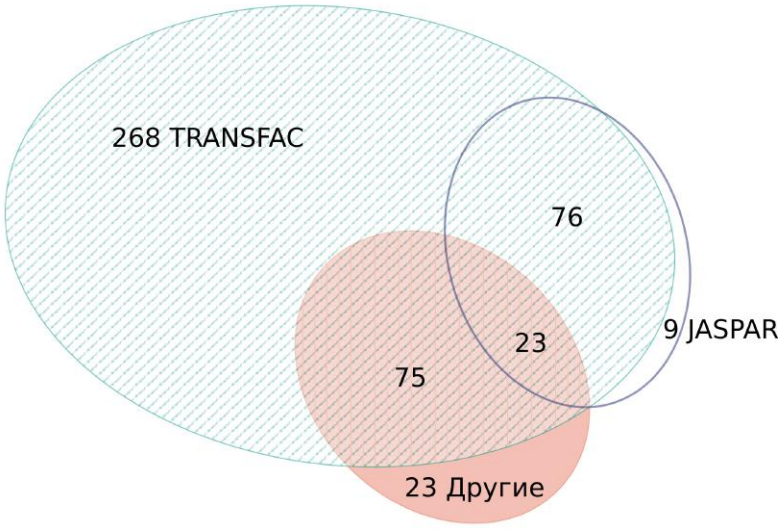


Рисунок 25. Диаграмма Венна, иллюстрирующая пересечение исходных данных (по факторам транскрипции) между основными источниками, использованными при построении базового релиза коллекции мотивов HOCOMOCO.

5.1.1.3. Вычислительная идентификация мотивов

Универсальные параметры для определения мотивов подобрать трудно, поскольку источники данных разнородны по объему и достоверности. Было принято решение использовать ChIPMunk в нескольких режимах и затем провести экспертное курирование результатов. Использовалось 4 режима: поиск мотива «сверху-вниз» и «снизу-вверх» (от максимальных длин к минимальным и от минимальных длин к максимальным в диапазонах от 7 до 25 п.н.), а также с априорными ограничениями на «форму» самого мотива: однобоксового либо двухбоксового (вписанного в один или два витка спирали ДНК по аналогии с программой HeliCis [Larsson, Lindahl, Mostad, 2007], см. также «Материалы и методы»).

5.1.1.4. Экспертное курирование результатов

В ходе экспертного курирования для каждого фактора транскрипции мы выделяли наиболее релевантную модель среди четырех, построенных ChIPMunk. Одновременно для каждого мотива была сделана оценка качества, которая при необходимости позволяет выделять в коллекции наиболее достоверный поднабор мотивов ценой уменьшения разнообразия покрытых факторов транскрипции.

Достаточно часто результаты ChIPMunk в различных режимах отличались незначительно, например для сильных, четко выраженных паттернов, узнаваемых белками CTCF и REST. В таких случаях мы выбирали наиболее простой вариант (без априорной установки формы либо с однобоксовой формой). В случае, когда результаты в различных режимах заметно отличались, мы руководствовались суммарным весом выравнивания (которое отражает как число включенных в него последовательностей, так и их суммарную достоверность). Для небольших наборов последовательностей мы выбирали наиболее короткие мотивы, считая, что достоверно определить частоты букв во фланкирующих позициях по малой выборке невозможно. Наконец, в двух десятках случаев для одного фактора транскрипции были выбраны сразу две модели, если была экспериментально показана способность фактора связываться с ДНК в форме мономера и димера либо гомо- или гетеродимера.

Эмпирические критерии для оценки качества мотивов

При оценке качества мотивов мы руководствовались следующими соображениями:

(1) Достоверные мотивы связывания имеют характерное распределение информационного содержания по колонкам: при лого-визуализации хорошо выделяется кор мотива и фланкирующие позиции, значимость которых падает при удалении от кора.

(2) Достоверные мотивы стабильно извлекаются вне зависимости от конкретных настроек программ идентификации мотивов; в случае ChIPMunk мы ожидали, что достоверные мотивы будут иметь сравнимую длину, похожие консенсусные последовательности и число сайтов в выравнивании, вне зависимости от режима запуска.

(3) Достоверные мотивы демонстрируют сходство с известными паттернами, узнаваемыми как самим фактором транскрипции, так и членами того же структурного семейства.

(4) Большое число сайтов связывания в выравнивании (сотня и более) также косвенно говорит в пользу достоверности мотива, в частности, позволяет определить частоты нуклеотидов в слабых, фланкирующих позициях. Однако большое число последовательностей не гарантирует хорошего выравнивания и, чаще всего, приходит из высокопроизводительных экспериментов со своими особенностями и систематическими ошибками. То есть, количественный критерий имеет смысл, но не может быть определяющим.

Рейтинги качества мотивов HOCOMOCO

В первой сборке HOCOMOCO мы использовали 6 рейтингов качества от A (наилучший) до F (неуспех, failure). Рейтинги качества от A до D присваивались мотивам известных факторов транскрипции (включая белки из базы TcoF-DB [Schaefer, Schmeier, Bajic, 2011]), и другим белкам, для которых был найден достоверный мотив связывания, и по литературным данным были свидетельства в пользу функциональной роли белка как фактора транскрипции (использовались, в том числе, результаты экспертного курирования факторов транскрипции, позднее опубликованные в ходе проекта FANTOM5 [Forrest и др., 2014]). Качество A присваивалось мотивам, удовлетворявшим всем четырем критериям достоверности (см. выше). Качество B присваивалось мотивам, построенным на больших наборах последовательностей, которые удовлетворяли как минимум двум из оставшихся

условий. Качество C присваивалось мотивам, построенным на ограниченных выборках, но удовлетворяющим трем прочим условиям. Мотивы качества D чаще всего представляли только часть известного консенсуса или вообще не имели хорошо выраженного корового участка с высоким информационным содержанием. Качество E (error, ошибка) и F присваивалось моделям, для которых не было уверенности в том, что соответствующий белок действительно является фактором транскрипции и способен специфически узнавать ДНК; в таких случаях консенсус или отсутствовал (получалось выравнивание без выраженных консервативных позиций), или соответствовал известному фактору транскрипции другого, нерелевантного семейства.

5.1.1.5. Обзор первого релиза коллекции

Первый публичный релиз v9 коллекции HOCOMOCO (HOmo sapiens COmprehensive MOtif COllection) содержал 426 мотивов качества A-D для 401 фактора транскрипции человека (среди них 52/87/139 мотивов наилучшего качества A/B/C). Средняя длина и медиана длин мотива составляла 12 п.н. (что сравнимо с одним витком спирали ДНК), модели качества A и B были построены по наборам последовательностей, в среднем интегрирующим два и более источника данных.

С использованием ограниченного набора ChIP-Seq пиков ENCODE для нескольких десятков факторов транскрипции мы провели сравнительное тестирование качества распознавания мотивов из TRANSFAC, JASPAR и HOCOMOCO путем оценки площади под ROC-кривыми, см. **Рисунок 26**. Построенные модели HOCOMOCO оказались более точными в 90% случаев. Позднее, достоверность мотивов, представленных в HOCOMOCO v9, была подтверждена и результатами независимых тестирований на других наборах факторов [Dabrowski и др., 2015; Kibet, Machanick, 2015].

Чтобы изучить иерархию мотивов в коллекции мы провели кластеризацию[38] мотивов A-C качества (использовался метод невзвешенного попарного среднего, Unweighted Pair-Group Method Using Arithmetic Averages, UPGMA) по сходству Жаккара, оцененному с помощью MACRO-APE для *P*-значений 0.0005. Кластеры

[38] Иерархическое дерево HOCOMOCO v9, визуализация jsPhyloSVG [Smits, Ouverney, 2010] http://autosome.ru/HOCOMOCOS/addon/hocomoco_clusters.html

были выделены путем сбора ПВМ, оказавшихся на одной ветви, останавливая траверс дерева в точке, когда минимальное попарное сходство между членами кластера становилось меньше 0.05. Дерево, визуализированное с помощью jsPhyloSVG [Smits, Ouverney, 2010] показано на **Рисунке 27**, где можно отметить, что ряд характерных семейств мотивов был успешно объединен в кластеры.

На момент публикации HOCOMOCO v9 представляла собой наиболее репрезентативную базу данных мотивов факторов транскрипции человека (поскольку для многих белков известны прямые гомологи у мыши, опционально мы предоставляли и картирование идентификаторов мотивов на факторы транскрипции мыши по базе UniProt).

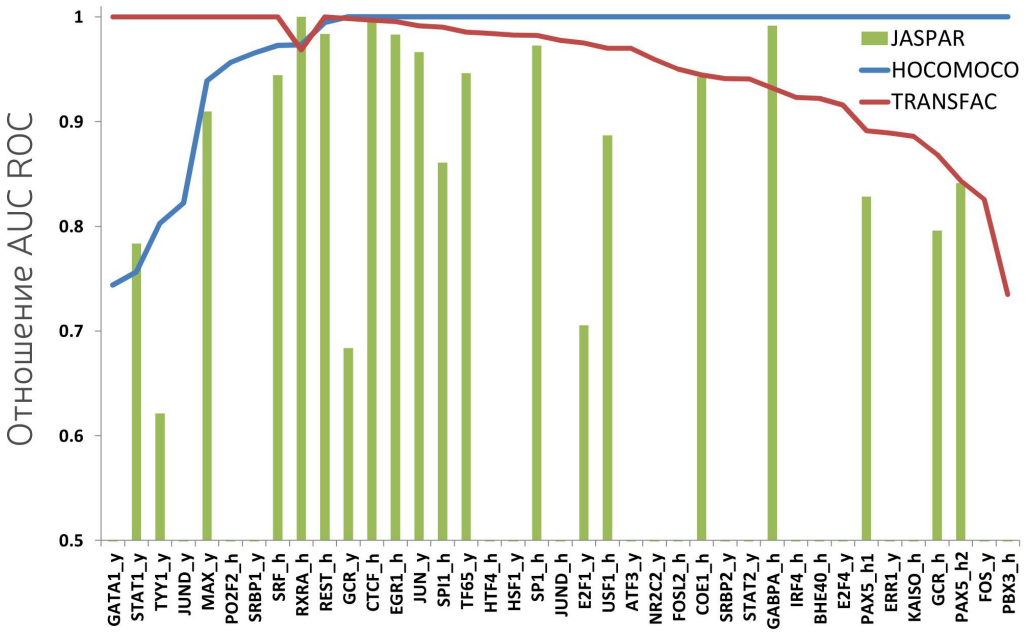


Рисунок 26. Сравнение относительных площадей (нормированы на наибольшую) под ROC-кривыми для моделей JASPAR (зеленые столбики), TRANSFAC (красная кривая) и HOCOMOCO (синяя кривая).

Единица соответствует наилучшей модели с наивысшим значением AUC ROC для конкретного фактора транскрипции. Точки по оси X соответствуют контрольным выборкам (не использованным при построении мотивов) для различных факторов транскрипции. В случае, когда в коллекции присутствовало более 1 мотива для конкретного фактора транскрипции, выбирался наилучший. Рисунок адаптирован из работы [Kulakovskiy и др., 2013a].

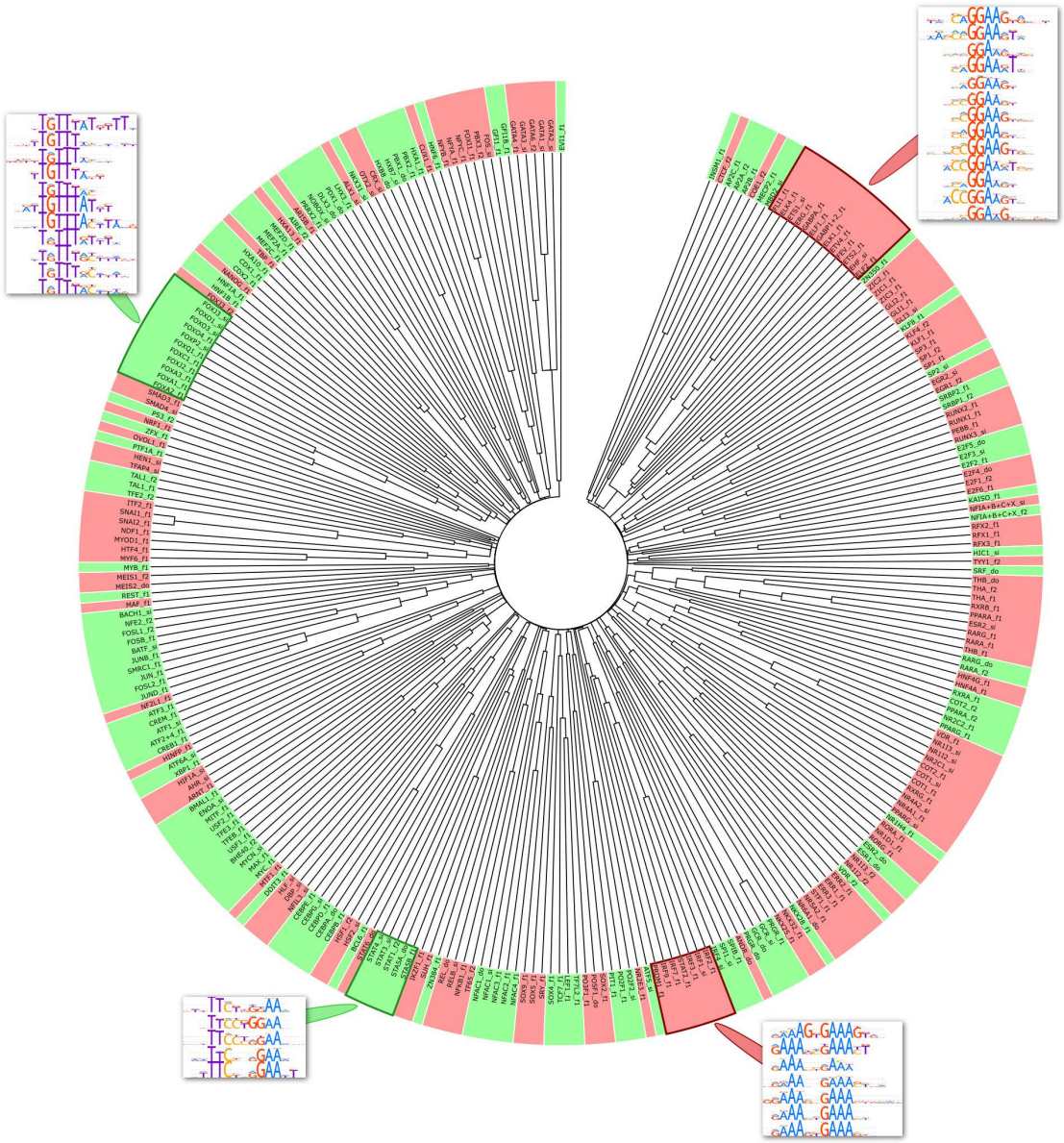


Рисунок 27. Круговая дендрограмма, иллюстрирующая иерархию достоверных мотивов из HOCOMOCO.

Кластеры мотивов для удобства показаны чередующимися цветами. Примеры скластеризованных мотивов, соответствующие известным семействам факторов транскрипции (ETS, STAT, IRF, FOX), показаны группами лого-диаграмм. Рисунок адаптирован из работы [Vorontsov, Kulakovskiy, Makeev, 2013].

5.1.2. Расширение коллекции путем систематического анализа данных ChIP-Seq

Основной вклад в первый публичный релиз коллекции HOCOMOCO внесли «догеномные» данные, но наиболее удачные модели были построены на основе данных ChIP-Seq. Возникла естественная мотивация расширить коллекцию с упором на данные высокопроизводительного секвенирования и провести полномасштабное сравнительное тестирование-«бенчмарк» позиционно-весовых матриц для сайтов разнообразных факторов транскрипции (в первой публичной версии тестирование было проведено лишь для пары десятков факторов транскрипции, для которых были доступны стартовые данные ENCODE).

Систематическое использование ChIP-Seq стало возможным благодаря партнерам из Института системной биологии (Новосибирск), которые переработали результаты нескольких тысяч опубликованных экспериментов ChIP-Seq для факторов транскрипции мыши и человека и представили результаты в базе данных GTRD[39].

Кроме того, для нескольких сотен факторов транскрипции были опубликованы данные HT-SELEX (*in vitro* [Jolma и др., 2013]), и было интересно оценить степень их достоверности и применимости для практического распознавания сайтов в геноме, которые связывают белок *in vivo* по данным ChIP-Seq.

Объемы ChIP-Seq и HT-SELEX данных достаточны для систематического построения расширенных моделей, потенциально превосходящих классические позиционно-весовые матрицы, а наличие нескольких независимых экспериментов для конкретного фактора транскрипции позволяет улучшить надежность тестирования. Мы приняли решение дополнительно построить динуклеотидные позиционно-весовые матрицы на основе данных HT-SELEX и ChIP-Seq и включить такие модели в общее сравнительное тестирование.

В предыдущей версии мы ограничились картированием идентификаторов моделей на белки мыши. Но в GTRD были доступны прямые данные ChIP-Seq для сотен факторов транскрипции мыши, и это позволило выделить для них отдельную группу моделей.

---

[39] Gene Transcription Regulation Database. http://gtrd.biouml.org/

5.1.2.1. Схема построения обновленной коллекции

Общая идея построения коллекции сохранилась: мы используем ChIPMunk для идентификации мотивов, курируем результаты на предмет согласованности мотивов с уже известными в рамках одного семейства, и проводим тестирование качества распознавания сайтов связывания в данных ChIP-Seq, не использованных при построении модели. В результате, мы выбираем один наилучший базовый мотив, хорошо описывающий общий паттерн сайтов связывания фактора транскрипции (не фокусируясь на особенностях конкретных экспериментов или типах клеток). Схема анализа представлена на **Рисунке 28**.

Анализ данных ChIP-Seq

В качестве источника единообразных данных ChIP-Seq выступала база GTRD (релиз сентября 2013 года): чтения картированы Bowtie [Langmead и др., 2009], пики выделены с помощью SISSRS [Jothi и др., 2008]. Использованный релиз содержал результаты 1690 экспериментов и покрывал 392 (96) фактора транскрипции для человека (мыши), учитывая только эксперименты с 200 и более идентифицированными пиками. Для каждого эксперимента пики были ранжированы по высоте (т.е. по максимальному покрытию чтениями) и 1000 наилучших были отобраны из каждого набора; в 652 наборах присутствовало менее 1000 пиков и в анализ были взяты все доступные.

Пики четных рангов использовались для идентификации мотивов с помощью ChIPMunk и diChIPMunk в диапазоне длин от 22 до 11 п.н. (режим «одно либо ни одного вхождения в последовательность»), явным образом учитывалась информация о форме пика в форме профилей покрытия. В каждом наборе данных итеративно идентифицировались два мотива (предполагая, что для конкретного эксперимента наиболее консервативный мотив может покрывать только небольшую часть выборки и отражать статистические артефакты, например повторы или неотфильтрованные ПЦР-дубликаты чтений, либо соответствовать типичному кофактору). Для каждого набора пиков в дальнейший анализ был взят мотив, покрывающий их большую часть. Пики нечетных рангов использовались как независимый контроль при тестировании.

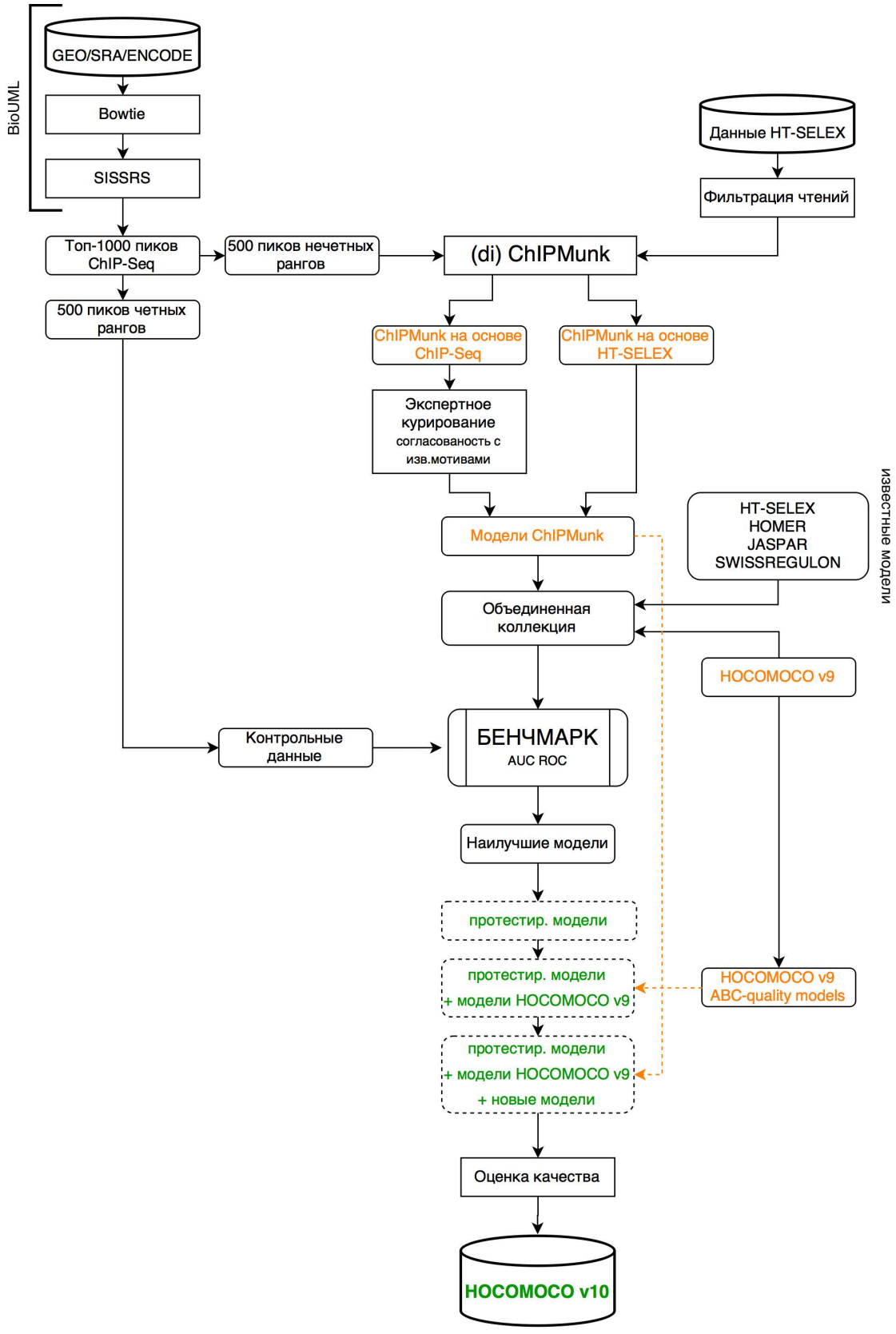


Рисунок 28. Схема построения HOCOMOCO v10.

Рисунок адаптирован из работы [Kulakovskiy и др., 2016].

Предварительное экспертное курирование результатов идентификации мотивов

Все мотивы, идентифицированные в ChIP-Seq данных, были вручную проаннотированы. Для дальнейшего анализа и сравнительного тестирования были взяты модели, удовлетворяющие следующим критериям: (1) похожие на любой из известных мотивов среди опубликованных результатов HT-SELEX и предыдущей сборки HOCOMOCO (как минимум 5% сходства по Жаккару); для непрошедших порог сходства по Жаккару дополнительно проверяли сходство консенсусов; (2) согласующиеся с мотивами факторов того же семейства либо (3) как минимум, с хорошо выраженным консенсусом (основываясь на анализе лого-диаграмм).

Характерный пример, не прошедший этап курирования, это мотив, идентифицированный в пиках ChIP-Seq фактора ARID3A. Идентифицированный паттерн однозначно соответствовал мотиву связывания FOX-семейства. С одной стороны, это не является техническим артефактом идентификации мотивов (и подтверждается независимыми анализом FactorBook[40]). С другой стороны это не согласуется ни с информацией о структурных семействах, ни между ChIP-Seq для ARID3A в различных клеточных линиях. Таким образом, идентифицированный мотив (несмотря на статистическую обогащенность в исследованном наборе данных и выраженный консенсус) не прошел этап курирования.

В сумме лишь около 50% мотивов прошли курирование и участвовали в автоматизированном сравнительном тестировании (692 для человека плюс 177 для мыши из полного множества 1690 наборов пиков). Это примерно соответствует имеющимся в литературе оценкам средней надежности опубликованных данных ChIP-Seq [Marinov и др., 2014] и согласуется с последующей оценкой качества использованных наборов ChIP-Seq данных (обсуждается ниже в разделе, посвященном сравнительному тестированию мотивов).

Обработка данных HT-SELEX

Исходные чтения, полученные в результате анализа связывания 542 факторов транскрипции с помощью технологии HT-SELEX, были переобработаны. Недостоверно прочитанные сегменты последовательностей были замаскированы

[40] ARID3A – FactorBook
http://www.factorbook.org/human/chipseq/tf/ARID3A

поли**N** (мы контролировали как минимум среднюю оценку Phred [Ewing и др., 1998] как минимум на уровне Q30 в скользящем окне 10 п.н.). Затем подсчитывали точное число прочтений каждого олигонуклеотида для всех соседних пар (*n*, *n*+1) SELEX-циклов и выбрали олигонуклеотиды, обогащенные при переходе от предыдущего цикла к последующему хотя бы для одного *n*. Наибольшее среди всех циклов число прочтений было использовано в качестве веса каждой последовательности; идентификация мотивов выполнялась с помощью ChIPMunk и diChIPMunk раздельно для каждого эксперимента (мотивы HT-SELEX-R) и, дополнительно, интегрируя данных всех экспериментов (для факторов транскрипции с несколькими экспериментами, мотивы HT-SELEX-I).

5.1.2.2. Коллекции мотивов, использованные в сравнительном тестировании

Для проведения сравнительного тестирования мы использовали несколько публичных коллекций мотивов: JASPAR, HOMER, SwissRegulon, исходные мотивы, определенные авторами HT-SELEX, и предыдущую сборку HOCOMOCO, дополненную моделями для регуляторов плюрипотентности [Papatsenko и др., 2015].

Таблица 3. Коллекции мотивов, использованные на этапе сравнительного тестирования. Построенные *de novo* наборы моделей выделены цветом. В тестирование вошли факторы транскрипции, для которых были доступны ChIP-Seq данные.

Коллекция мотивов	**Число факторов транскрипции (идентификаторов UniProt) для человека (мыши)**	**Число факторов транскрипции, мотивы связывания которых участвовали в тестировании**
ChIP-Seq моноПВМ	120 (64)	114 (63)
HT-SELEX-R моноПВМ	400 (2)	50
HT-SELEX-I моноПВМ	119	21
HOCOMOCO v9	405 (393)	102 (51)
JASPAR	130 (68)	75 (34)
SwissRegulon	337	71
HOMER	123	62
HT-SELEX	404 (2)	53
ChIP-Seq диПВМ	120 (64)	115 (63)
HT-SELEX-R диПВМ	400 (2)	52
HT-SELEX-I диПВМ	119	21

Сравнительное тестирование проводилось только на наборах пиков, которые успешно прошли стадию фильтрации (см. ниже). Были успешно протестированы почти все новые мотивы, определенные в ChIP-Seq. Другие коллекции были хуже покрыты тестированием, доля протестированных мотивов составляла от 25% (HOCOMOCO v9) до 50% (JASPAR). Список факторов транскрипции, для которых были протестированы мотивы, перекрывался между коллекциями; полный объединенный набор составлял 127 (69) факторов для человека (мыши). В **Таблице 3** новые мононуклеотидные модели (из ChIP-Seq и HT-SELEX данных) выделены зеленым фоном, а новые динуклеотидные матрицы – оранжевым. Как и в предыдущем релизе HOCOMOCO v9 для удобства мы сопоставили все факторы транскрипции идентификаторам по базе белков UniProt [The UniProt Consortium, 2012].

5.1.2.3. Организация сравнительного тестирования

Чтобы сравнить точность новых и существующих моделей мы использовали AUC ROC, взяв за образец и расширив методологию предыдущей версии HOCOMOCO. Для каждого фактора транскрипции мы тестировали все имеющиеся мотивы на всех имевшихся наборах пиков для этого фактора. На разных порогах оценок весовых матриц мы подсчитывали долю истинных положительных предсказаний (долю пиков с как минимум одним вхождением мотива) и ожидаемую долю ложных положительных предсказаний среди случайных последовательностей сравнимой длины (см. «Материалы и методы»). По сравнению с предыдущей версией HOCOMOCO мы включили в анализ динуклеотидные матрицы и единообразно учитывали динуклеоитдный состав пиков при подсчете *P*-значений и моно- и динуклеотидных мотивов мотивов для оценки ложноположительных предсказаний.

Все наборы пиков ChIP-Seq в GTRD вычислительно построены одинаковым образом, но это не гарантирует сравнимого качества исходных данных и сравнимой обогащенности вхождениями мотивов (в силу биологических или технических причин). В то же время, не все из новых и известных мотивов действительно отражают реальные особенности ДНК-белкового взаимодействия. То есть, необходимо одновременно выявить как неудачные эксперименты ChIP-Seq, так и неверные мотивы.

Пусть для заданного фактора есть *N* мотивов связывания. Для каждого набора пиков подсчитаем взвешенную площадь под кривой ROC:

$$w\text{AUC}_{\text{набор}} = \frac{\sum_{\text{мотивы}} \text{AUC}(\text{мотив},\text{набор})}{N}$$

Аналогичное значение подсчитаем для каждого мотива:

$$w\text{AUC}_{\text{мотив}} = \frac{\sum_{\text{наборы}} \text{AUC}(\text{мотив},\text{набор}) \cdot w\text{AUC}_{\text{набор}}}{\sum_{\text{наборы}} w\text{AUC}_{\text{набор}}}$$

Эта процедура позволяет использовать информацию из нескольких наборов пиков чтобы оценить качество мотива и, обратно, оценить качество каждого набора пиков на основании всех имеющихся мотивов (понимая качество набора в смысле обогащенности сайтами связывания). Использование взвешенного среднего для AUC ROC позволяет лучшим мотивам и наборам пиков вносить больший вклад в оценку всех остальных.

Тем не менее, это не полностью решает проблему ошибочных мотивов и наборов пиков без вхождений мотива (либо ошибочных, либо отражающих сложный специальный случай: и то и другое не позволяет использовать эти пики для тестирования базовой универсальной модели мотива). Чтобы избежать систематических искажений от ошибочных мотивов и наборов пиков, для каждого фактора транскрипции мы итеративно убирали из общей выборки наборы пиков с *w*AUC < 0.65 и мотивы с *w*AUC < 0.65 и пересчитывали значения *w*AUC. Пороговое значение для *w*AUC было выбрано из простых соображений: модели HOCOMOCO v9 качества A и B должны оставаться в сравнительном тестировании (поскольку ранее успешно прошли наиболее строгую процедуру курирования).

В результате, 786 (206) наборов пиков для человека (мыши) прошли процедуру фильтрации по *w*AUC (см. также **Таблицу 3** выше с обзором по числу факторов транскрипции). Для каждого фактора транскрипции мы смогли выбрать наилучший мотив, основываясь на соответствующем наилучшем значении *w*AUC.

5.1.2.4. Сборка итоговой коллекции

После проведения сравнительного тестирования напрашивается полностью автоматическая процедура сборки коллекции из мотивов с наилучшим *w*AUC. Однако, есть два конфликтующих соображения. Во-первых, единообразный

вычислительный протокол идентификации мотивов (в нашем случае – с помощью ChIPMunk) является одним из удобных свойств коллекции (можно рассчитывать, что различия мотивов не являются техническими следствиями применения различных алгоритмов). Во-вторых, с точки зрения практических приложений часто удобно иметь основную коллекцию в виде классических мононуклеотидных весовых матриц (даже если они проигрывают динуклеотидным по качеству распознавания сайтов связывания).

Стратегией по умолчанию все равно остается выбор наилучшего мотива по результатам тестирования AUC ROC. Новые модели на основе данных ChIP-Seq оказались наилучшими в абсолютном большинстве случаев. Выбор был сложнее если данных ChIP-Seq и, соответственно, результатов тестирования, не было для конкретного фактора транскрипции (например, ни один набор пиков не прошел этапа фильтрации по *w*AUC). В этом случае мы, в первую очередь, включали в новую коллекцию модели наилучшего качества ABC из HOCOMOCO v9, а модели HT-SELEX-I/R (как построенные только по данным *in vitro*) и D-качества из HOCOMOCO v9 были следующими кандидатами. Наконец, для случаев, когда динуклеотидные матрицы показывали wAUC лучше, чем у мононуклеотидных, соответствующие динуклеотидные мотивы отбирались в дополнительную коллекцию «уточненных» моделей.

Для моделей, участвовавших в тестировании, был создан единый рейтинг качества, основанный на AUC ROC и следующих соображениях: (1) модели наивысшего качества показывают AUC ROC выше оптимального порога как минимум на двух независимых наборах пиков; (2) если доступен только один набор пиков ChIP-Seq, он является достоверной валидацией для моделей, не основанных на тех же ChIP-Seq данных; (3) для моделей, не участвовавших в тестировании, может быть использован оригинальный рейтинг качества HOCOMOCO. Формализованное представление алгоритма для присвоения мотивам рейтингов качества представлено на **Рисунке 29** в виде блок-схемы.

К сожалению, формализация не полностью избавляет от эмпирических параметров; так, выбор оптимального порога AUC для моделей наивысшего качества сделан на основе курированных метрик качества HOCOMOCO v9 в качестве образца. Оптимальный AUC в 0.8 (напомним, при максимуме в 1 и значении 0.5

соответствующим случайной классификации) достигался для 70% (50 из 82) моделей качества AB из HOCOMOCO v9 и в два раза меньшей долей моделей качества CD (35%, 13 из 36 прошедших тестирование).

Для ряда белков возможно связывание в форме мономера и димера. Мы сохраняли информацию о вторичных (secondary) или однобоксовых (single-box) мотивах под качеством S, чтобы явно выделить их в новой коллекции. Короткие вторичные мотивы редко могут достичь высоких AUC в силу малых длин и, как следствие, ограниченной «дискриминирующей способности», но содержат важную информацию, которая может пригодиться для специфических задач, например, в анализе композитных элементов. Модели качества S были включены для 40 (31) факторов человека (мыши), основываясь на существующей курированной аннотации HOCOMOCO v9, либо, когда новый мотив HOCOMOCO v10, выбранный по результатам сравнительного тестирования, соответствовал заметно более длинному или отличающемуся паттерну (например, двухбоксовому).

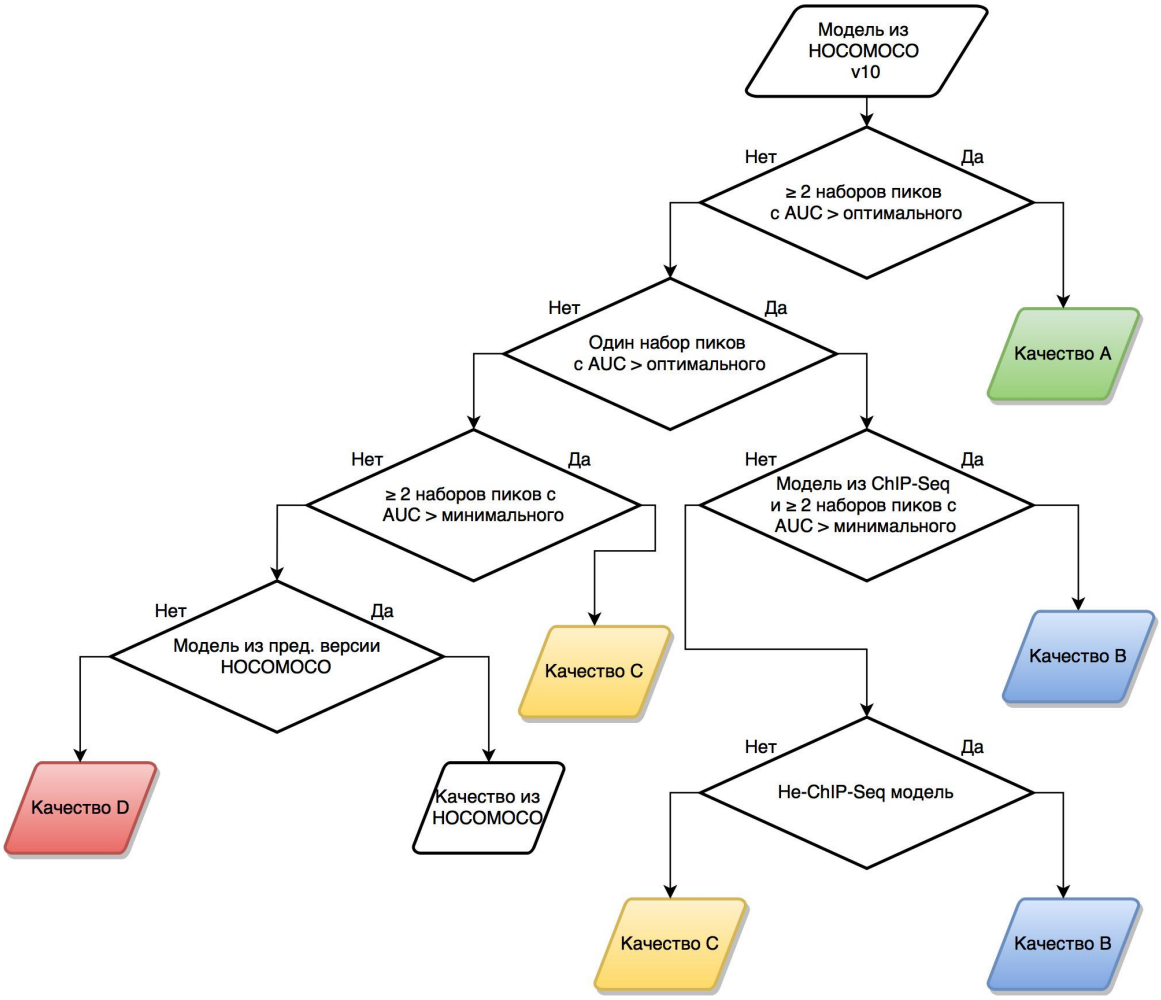


Рисунок 29. Схема присвоения качества моделям HOCOMOCO v10. Рисунок адаптирован из работы [Kulakovskiy и др., 2016].

5.1.2.5. Обзор итоговой коллекции

В HOCOMOCO v10 значительно расширен состав мотивов для факторов транскрипции человека и мыши. Были использованы 992 наиболее достоверных из 1690 исходных наборов пиков ChIP-Seq и результаты 542 экспериментов HT-SELEX. На момент публикации HOCOMOCO v10 была наиболее полной систематической коллекцией мотивов, содержащей курированные модели, которые по возможности прошли верификацию на независимых данных *in vivo*.

Таблица 4. Число факторов транскрипции человека, покрытых известными мотивами, представленными в различных коллекциях.

Коллекция	**Число факторов транскрипции человека**
HOMER	123
JASPAR (CORE VERTEBRATE)	130
SwissRegulon	337
HT-SELEX *(исходные модели)*	404
HOCOMOCO v9	401
HOCOMOCO v10	600

Данные соответствуют релизам коллекций, доступным на момент публикации результатов проекта FANTOM5 [Forrest и др., 2014].

По сравнению с предыдущим релизом, HOCOMOCO v10 покрывает мотивами еще две сотни факторов транскрипции человека, и, дополнительно, содержит более сотни динуклеотидных моделей.

Суммируя: HOCOMOCO v10 содержит 600 (395) мотивов связывания факторов транскрипции человека (мыши), 273 (262) наиболее достоверных моделей высокого качества (ABC, курированных в HOCOMOCO v9 либо протестированных в ходе обновления), и дополнительно 86 (52) динуклеотидных моделей (включены только модели, превосходящие по точности соответствующие мононуклеотидные аналоги).

Мотивы для 92 (52) факторов транскрипции человека (мыши) были идентифицированы в пиках ChIP-Seq, мотивы для 193 (1) факторов транскрипции были выявлены в данных HT-SELEX, 315 (342) мотива были унаследованы из HOCOMOCO v9. Кроме того, для 40 (30) факторов транскрипции HOCOMOCO v10 включает вторичные мотивы (в основном, унаследованные из HOCOMOCO v9).

Результаты сравнительного тестирования коллекций

Для сравнительной оценки общего качества мотивов в различных коллекциях для каждой коллекции мы подсчитали число факторов транскрипции, мотивы которых показали наилучший *w*AUC среди всех коллекций.

Для начала каждая коллекция весовых матриц сравнивалась с объединением всех остальных коллекций (за вычетом итоговой коллекции HOCOMOCO v10 и динуклеотидных мотивов). Мотивы JASPAR и HOCOMOCO v9 оказались наиболее достоверными для пары десятков факторов транскрипции. Напротив, новые мотивы, идентифицированные в ChIP-Seq, были наилучшими для значительно более широкого спектра факторов. При независимом сравнении мотивов «только из HT-SELEX», примерно в половине случаев идентифицированные *de novo* мотивы превосходили оригинальные, но в общем тестировании объединенные мотивы на основе HT-SELEX показали посредственные результаты (не смотря на выбор наилучшего мотива среди всего множества HT-SELEX-моделей для каждого фактора транскрипции, учитывая как исходные модели от авторов эксперимента, так и наши результаты переобработки). Интересно, что ограничения HT-SELEX моделей в распознавании сайтов *in vivo* была замечены и другими исследователями [Kähärä, Lähdesmäki, 2013], и в наших предыдущих работах [Papatsenko и др., 2015]. Что касается динуклеотидных мотивов, в тех случаях, когда их удавалось построить, они оказывались не хуже либо превосходили известные мононуклеотидные по качеству распознавания сайтов связывания. Сводка результатов представлена на **Рисунке 30**.

В базе HOCOMOCO v10 мы сохранили значения *w*AUC и наилучшего достигнутого AUC в дополнительных полях аннотации 145 (76) мононуклеотидных моделей для факторов транскрипции человека (мыши). Это должно помочь исследователям оценивать практическую применимость конкретных мотивов.

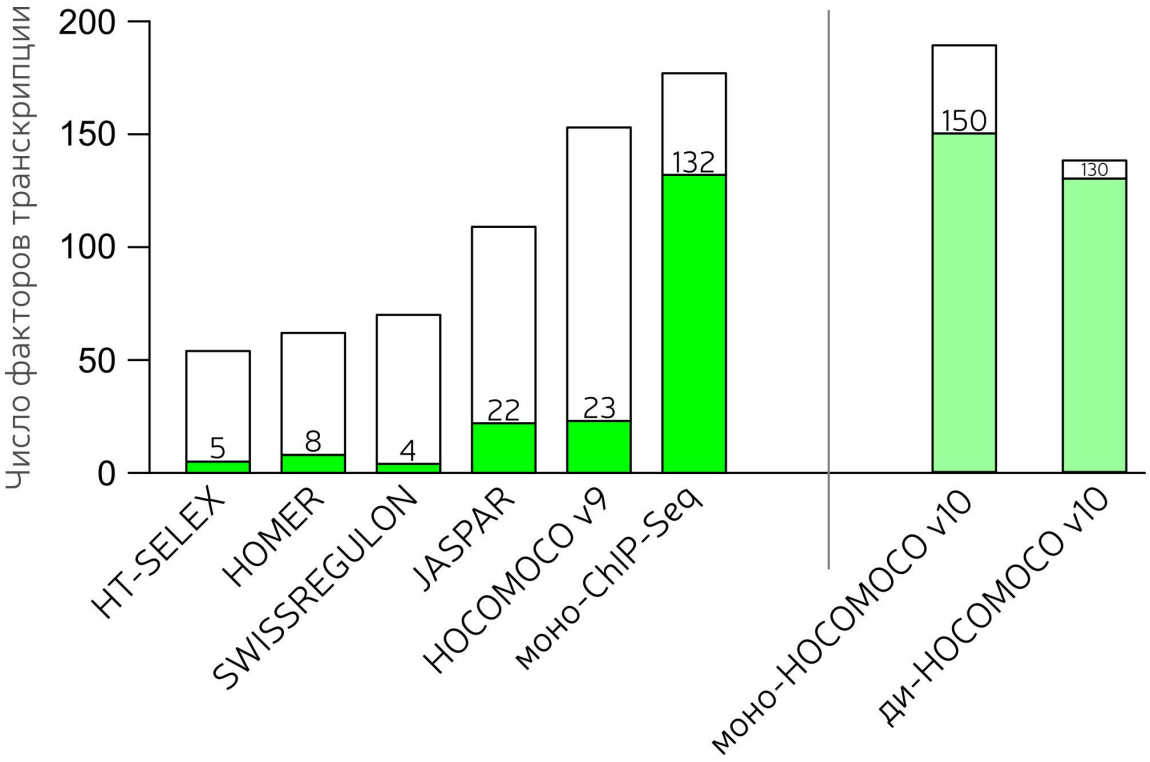


Рисунок 30. Результаты тестирования качества мотивов в различных коллекциях на основе данных ChIP-Seq для факторов транскрипции мыши и человека.

Полная высота полос показывает полное число факторов транскрипции, мотивы для которых были протестированы в конкретном случае. Зеленая часть полос соответствует числу факторов транскрипции, мотивы которых оказались наилучшими (имели наибольшие значения *w*AUC). Белая часть полосы для мотивов моно-HOCOMOCO v10 соответствует факторам транскрипции, для которых наилучшие модели не вошли в коллекцию (т.е. принадлежали коллекциям HOMER, SWISSREGULON, JASPAR и опубликованным ранее моделям HT-SELEX). Рисунок адаптирован из работы [Kulakovskiy и др., 2016].

5.1.2.6. Обсуждение результатов построения коллекции

Только 992 из 1690 наборов пиков ChIP-Seq прошли стадию фильтрации. Это не значит, что остальные наборы обязательно имеют низкое качество. Для оценки наборов пиков мы пользовались известными мотивами, и наборы пиков, содержащие принципиально новые мотивы для конкретных факторов транскрипции, оказывались в невыгодном положении. Для факторов транскрипции, связывающих характерные подтипы мотивов, в «смешанном» наборе пиков высокие значения AUC ROC могут быть фактически недостижимыми для единой модели. Кроме того, общий алгоритм обработки всех экспериментальных данных является практичным для серийного анализа, но может не учитывать какие-то специфические особенности конкретных экспериментов. Таким образом, мы не можем утверждать, что ChIP-Seq наборы были отфильтрованы полностью корректно: информация об узко-специфичных мотивах и ChIP-Seq-данных была потеряна «по построению». Также мы не сравнивали геномную локализацию сайтов связывания в различных наборах пиков, считая, что достаточно пронаблюдать обогащение вхождениями мотива; альтернативный подход может заключаться в поиске универсальных мотивов в специально отобранных пиках, воспроизводимых в различных экспериментах.

В тестировании мы использовали тысячу наиболее высоких пиков из каждого набора, считая, что они наиболее достоверны в смысле реального связывания белка и вычислительной идентификации. Потенциально, наивысшие пики могут иметь статистический сдвиг в пользу какого-то подтипа сайтов связывания, но чаще всего нескольких сотен последовательностей достаточно для построения базовых моделей мотивов, например, как это сделано в FactorBook [Wang и др., 2013a]. Это позволяет игнорировать, но не решить проблему выбора оптимального поднабора пиков для идентификации мотива.

Использование пиков четных/нечетных рангов для идентификации мотива и тестирования, соответственно, на взгляд традиционного машинного обучения кажется менее корректным, чем случайное сэмплирование. Однако, строго говоря, случайное сэмплирование, обеспечивающее устойчивое выделение мотива, выполняет сам ChIPMunk. А использование фиксированных близких рангов пиков для построения мотива и валидации является логичным, учитывая, что достоверность сайтов связывания скоррелирована с высотой пиков. Кроме того, крайне невероятно,

чтобы сэмплирование «через один» вносило какой-либо биологически-обоснованный шум. Разумеется, в более слабых пиках могут существовать характерные подтипы или альтернативные мотивы, но достаточно трудно доказать, что они имеют прямой биологический смысл и не относятся к кофакторам или артефактам конкретного эксперимента.

Наша цель состояла в том, чтобы создать надежную базовую коллекцию, пригодную для различных практических приложений. HOCOMOCO v10 содержит модели для 600 факторов транскрипции, что составляет менее половины всех факторов транскрипции человека [Wingender и др., 2015]. При отсутствии прямых экспериментальных данных присвоить недостающие мотивы факторам транскрипции нетривиально, но возможно, например по гомологии ДНК-связывающих доменов, что было масштабно проделано на межвидовом уровне в работе [Weirauch и др., 2014].

В этом смысле интересно сравнение мотивов факторов транскрипции мыши и человека, полученных при анализе прямых ChIP-Seq данных. В большинстве случаев мотивы для факторов мыши очень похожи на мотивы соответствующих ортологов у человека. Однако, есть и нетривиальные примеры. Так, у мотивов STAT1 отличается ориентация боксов: тандемный повтор у мыши и, чаще всего, палиндром у человеческого фактора, см. **Рисунок 31**. Вероятно эти особенности отражают различные варианты димеризации и скорее вызваны не межвидовыми различиями, а различиями в типах клеток или конкретных экспериментов. Возможно, в этом случае гибкая модель, допускающая различную ориентацию боксов и переменный спэйсер, предоставила бы более точную информацию.

Что касается применения динуклеотидных моделей, сравнение моно- и динуклеотидных моделей показывает большое сходство консенсусов и лого-визуализаций. Чаще всего diChIPMunk подбирает более длинные фланкирующие последовательности, чем ChIPMunk. Разница в длине существенна для белков, узнающих GC-боксы, например семейств SP и E2F. Идентификация мотива извлекает известный коровый участок, и достаточно тяжелые G-богатые фланки. Трудно сказать, правда ли это описывает свойства мотива, либо просто соответствует факту локализации GC-боксов в районах с повышенным GC-составом (результаты идентификации мотивов нестабильны и зависят от заданного фонового

распределения и режима идентификации мотива). В HOCOMOCO v10 мы сохранили более короткие версии мотивов под качеством S (альтернативные однобоксовые), но сам вопрос требует дальнейшего изучения.

Детального анализа требуют и некоторые модели качества D. Например, мотив связывания YBX1 D-качества унаследован от HOCOMOCO v9, но переобработка экспериментов ChIP-Seq показывает, что YBX1 вообще не имеет выраженной специфичности узнавания ДНК [Dolfini, Mantovani, 2012]. Разобраться в таких нетривиальных случаях, видимо, удастся только с появлением дополнительных данных, как *in vivo*, так и *in vitro*.

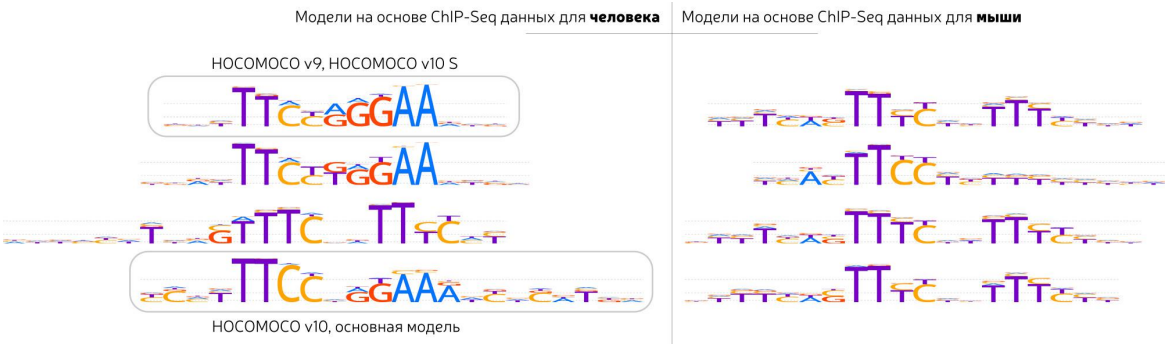


Рисунок 31. Мотивы связывания факторов транскрипции STAT1 мыши и человека, идентифицированные в различных набора данных ChIP-Seq.
Для сравнения показана модель из HOCOMOCO v9. Интересно, что один из экспериментов ChIP-Seq для человека показал подтип мотива, схожий с мышиным. Рисунок адаптирован из работы [Kulakovskiy и др., 2016].

5.1.3. Заключение по разделу

Коллекция HOCOMOCO доступна и в машинно-читаемом виде во множестве стандартных форматов файлов, и в человеко-читаемом виде через веб-интерфейс[41] и систему интерактивных фильтров. Основным идентификатором фактора транскрипции в HOCOMOCO является ключ по базе UniProt, но мы предоставляем ссылки и на другие ключевые базы данных (в том числе Entrez Gene[42], HGNC[43], MGI[44], FANTOM5 SSTAR[45]). В HOCOMOCO v10 мы явным образом включили информацию о структурных семейства ДНК-связывающих доменов согласно классификации TFClass[46]. На **Рисунке 32** представлено иерархическое дерево, показывающее подсемейства и семейства факторов транскрипции и их покрытие мотивами в HOCOMOCO v10. Интерактивная версия дерева на веб-сайте HOCOMOCO используется как стартовая навигационная страница. На сегодня публикации о HOCOMOCO собрали уже десятки цитирований, в том числе, HOCOMOCO как ключевой ресурс упомянута в фундаментальном обзоре по изучению геномных вариантов в сайтах связывания [Deplancke и др., 2016]. Мы надеемся, что наша коллекция мотивов будет и дальше полезна научному сообществу для исследований сайтов связывания в задачах регуляторной геномики.

[41] HOCOMOCO COmprehensive MOdel COllection. http://hocomoco.autosome.ru
[42] Home - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene
[43] HGNC database of human genes. http://www.genenames.org
[44] MGI - Mouse Genome Informatics. http://www.informatics.jax.org/
[45] FANTOM5_SSTAR. http://fantom.gsc.riken.jp/5/sstar/Main_Page
[46] TFClass: Classification of Human Transcription Factors and Mouse Orthologs. http://tfclass.bioinf.med.uni-goettingen.de

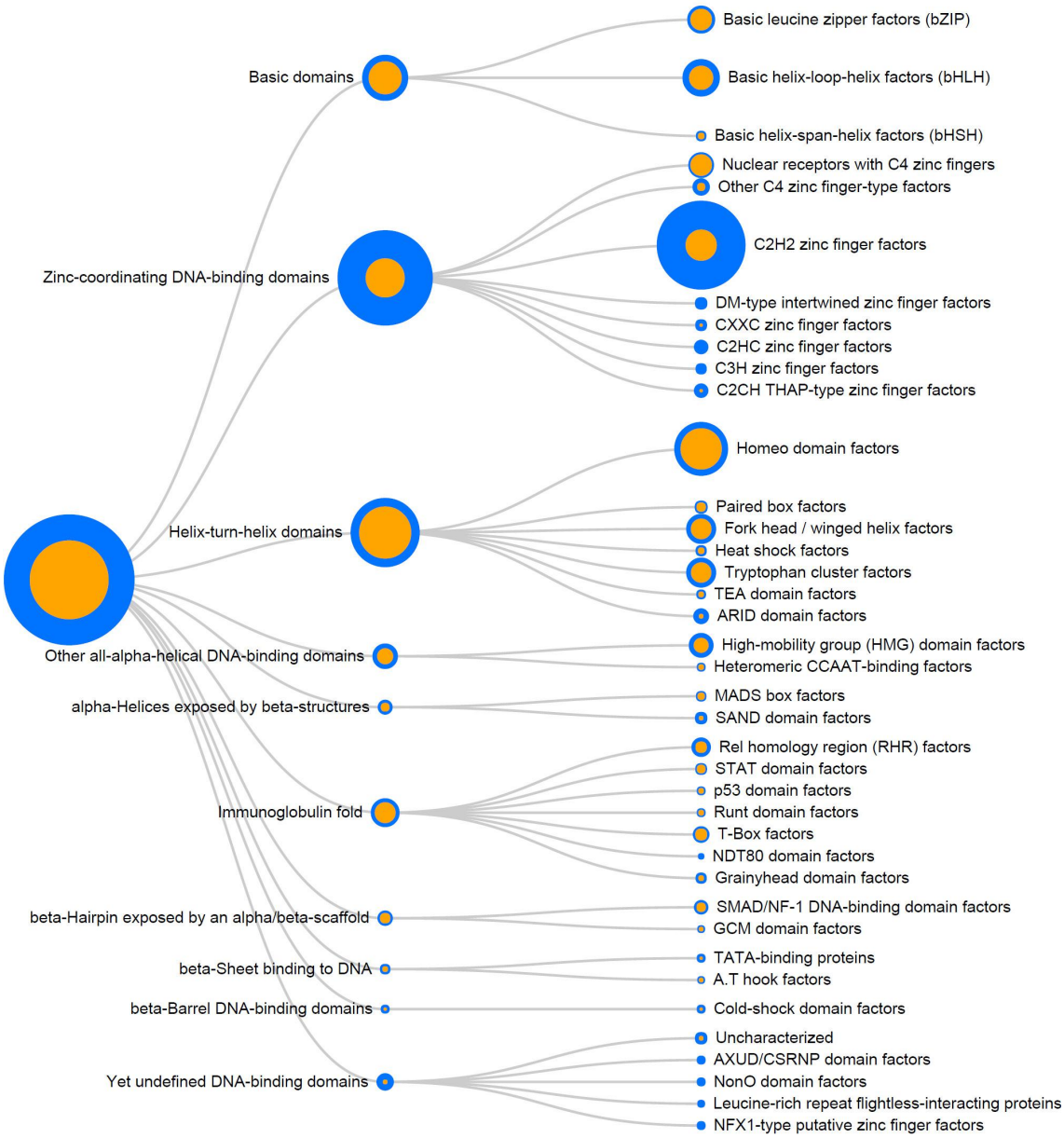


Рисунок 32. Покрытие основных структурных семейств факторов транскрипции моделями HOCOMOCO v10.

Площадь синих кругов пропорциональна полному числу членов конкретного семейства; оранжевые внутренние круги показывают долю факторов транскрипции, для которых доступны мотивы связывания. Классификация факторов транскрипции дана по TFClass [Wingender и др., 2015]. Рисунок адаптирован из работы [Kulakovskiy и др., 2016].

5.2. Практический анализ мотивов в избранных регуляторных системах

В этом разделе диссертации обсуждается применение авторских методов для анализа мотивов в конкретных задачах регуляторной геномики.

5.2.1. Мотивы и композитные элементы сайтов связывания факторов плюрипотентности OCT4/SOX2/NANOG

Одним из ключевых вопросов регенеративной биологии является устройство механизмов контроля самообновления клеток и плюрипотентности, и возможность их контролируемой настройки. Молекулярный анализ эмбриональных клеток и успехи в получении индуцированных стволовых клеток отдают ведущую роль в этих процессах факторам транскрипции. Среди них ключевыми считаются OCT4 (POU5F1), SOX2 и NANOG [Ivanova и др., 2006; Loh и др., 2006; Takahashi, Yamanaka, 2006]. Были предложены различные варианты генных сетей плюрипотентности, включающие заметно более длинный список регуляторов [Festuccia и др., 2013; Yang и др., 2010], но даже для трех основных факторов регуляторные взаимодействия остаются не до конца ясными. В работе [Papatsenko и др., 2015] был проведен высокопроизводительный анализ экспрессии ключевых генов в нескольких сотнях отдельных эмбриональных стволовых клеток мыши, использовался метод параллельной количественной ПЦР с помощью микрофлюидики. Анализ экспрессии генов позволил предположить наличие двух характерных субпопуляций клеток и реконструировать ключевой участок генной сети, потенциально обеспечивающей устойчивые альтернативные состояния. В частности, Дмитрием Папаценко было предложено существование некогерентной петли прямой-обратной связи (incoherent feedforward loops, iFFL [Goentoro и др., 2009]), включающей OCT4 и NANOG и предполагающей антагонизм этих двух ключевых регуляторов. Задачей анализа мотивов была независимая оценка, возможен ли антагонизм OCT4 и NANOG с точки зрения структуры регуляторных последовательностей. Для этого для OCT4, SOX2 и NANOG мы провели детальный анализ имеющихся в открытом доступе данных по связыванию ДНК *in vivo* и известных мотивов связывания.

5.2.1.1. Обзор доступных ChIP-Seq данных

Полногеномный профиль связывания факторов плюрипотентности в эмбриональных стволовых клетках мыши и человека исследовался с помощью метода ChIP-Seq в нескольких работах, включая проект ENCODE. Результаты части экспериментов были позднее единообразно обработаны Дж. Гоке (J. Göke), который любезно предоставил полученные данные.

Таблица 5. Источники данных ChIP-Seq для факторов плюрипотентности.

Идентификатор источника данных	**Биологический вид**	**Публикация**	**Факторы транскрипции**	**Релиз геномной сборки (по UCSC)**
CHEN2008	Мышь	[Chen и др., 2008]	NANOG, OCT4, SOX2	mm8
GOKECHEN	Человек	[Chen и др., 2008; Göke и др., 2011]	NANOG, OCT4, SOX2	mm9
ENCODE	Человек	[Dunham и др., 2012]	NANOG, OCT4	hg19
GOKE2011	Мышь, Человек	[Göke и др., 2011]	NANOG, OCT4, SOX2	mm9, hg19
CHIA2010	Человек	[Chia и др., 2010]	NANOG, OCT4	hg18

5.2.1.2. Схема вычислительного анализа

Схема анализа похожа на использованную при построении базы данных HOCOMOCO. Из ранжированных по достоверности списков мы извлекали 1000 наилучших пиков; пики нечетных рангов использовались для идентификации мотивов с помощью ChIPMunk, пики четных рангов использовались в качестве контрольных. Идентифицированные *de novo* мотивы сравнивались друг с другом и с известными мотивами с помощью ROC-кривых, контрольные наборы пиков выступали в качестве позитивной выборки. Схема анализа приведена на **Рисунке 33**.

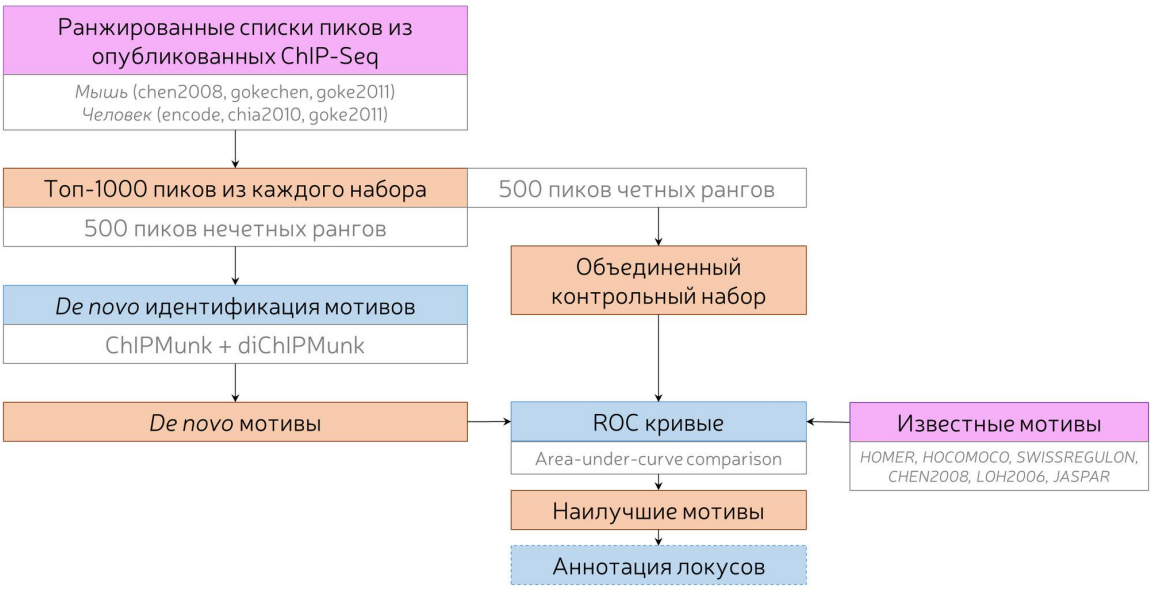


Рисунок 33. Схема систематической идентификации мотивов связывания для факторов OCT4, SOX2 и NANOG на основе ChIP-Seq данных.

Рисунок адаптирован из работы [Papatsenko и др., 2015].

5.2.1.3. Обзор известных мотивов связывания

Факторы плюрипотентности изучаются довольно активно, в публикациях представлена масса вариантов мотивов связывания, построенных с использованием различных типов экспериментальных данных. Одной из задач нашего анализа было определение наилучшей модели (в смысле распознавания *in vivo*-сайтов) среди известных и построенных *de novo*.

Таблица 6. Известные мотивы связывания ключевых факторов плюрипотентности.

Идентификатор мотива	**Источник**	**Число слов в выравн.**	**Публикация**
NANOG_CHEN2008	ChIP-Seq	80	[Chen и др., 2008]
NANOG_HOCOMOCO	HOCOMOCO v9 (по данн. TRANSFAC)	5	[Kulakovskiy и др., 2013a]
NANOG_HOMER	HOMER на базе ChIP-Seq	6781	[Heinz и др., 2010]
NANOG_LOH2006	ChIP-PET	100	[Loh и др., 2006]
NANOG_SWISSREGULON1	SwissRegulon	100	[Pachkov и др., 2013]
NANOG_SWISSREGULON2	SwissRegulon	50	[Pachkov и др., 2013]
OCT4_CHEN2008	ChIP-Seq	339	[Chen и др., 2008]
OCT4_HOCOMOCO	HOCOMOCO v9 (по данным JASPAR и TRANSFAC)	>1000	[Kulakovskiy и др., 2013a]
OCT4_HOMER	HOMER на базе ChIP-Seq	6798	[Heinz и др., 2010]
OCT4_JASPAR	JASPAR по данным ChIP-Seq [Chen и др., 2008]	1369	[Portales-Casamar и др., 2010]
OCT4_LOH2006	Loh2006 ChIP-PET	100	[Loh и др., 2006]
OCT4_SWISSREGULON1	SwissRegulon	50	[Pachkov и др., 2013]
OCT4_SWISSREGULON2	SwissRegulon	108	[Pachkov и др., 2013]
SOX2_CHEN2008	ChIP-Seq	260	[Chen и др., 2008]
SOX2_HOCOMOCO	HOCOMOCO v9 (по	>600	[Kulakovskiy и

	данным Jaspar и TRANSFAC)		др., 2013a]
SOX2_HOMER	HOMER на базе ChIP-Seq	1915	[Heinz и др., 2010]
SOX2_JASPAR	JASPAR по данным ChIP-Seq [Chen и др., 2008]	669	[Portales-Casamar и др., 2010]
SOX2_SWISSREGULON1	SwissRegulon	108	[Pachkov и др., 2013]
SOX2_SWISSREGULON2	SwissRegulon	50	[Pachkov и др., 2013]

Особое место занимают мотивы HT-SELEX [Jolma и др., 2013], построенные по массивным выборкам потенциальных сайтов, но обладающие вырожденными консенсусами. Для них результаты, полученные в ходе сравнительного тестирования, показывали ограниченное качество распознавания сайтов связывания (что согласуется с оценками, полученными при построении коллекции HOCOMOCO v10, см. соотв. раздел); в силу этого мы не включали эти мотивы в детальное сравнение. Выровненные лого-визуализации мотивов представлены на левых панелях на **Рисунке 34**. Всякий раз рамкой выделен бокс, связываемый самим фактором транскрипции (согласующийся между различными мотивами, в том числе, выявленными в *in vitro* данных).

5.2.1.4. Результаты идентификации мотивов *de novo* и сравнительного тестирования

Идентификация мотивов показала чрезвычайно стабильные результаты как при использовании различных источников данных, так и при сравнении «человек-мышь». Единственное исключение – тандемные повторы в качестве предпочтительных мотивов OCT4 у человека, идентифицированные в ChIP-Seq ([Chia и др., 2010] и [Göke и др., 2011]). Представленность повторов была ограничена поднабором наилучших пиков из первой тысячи, вопрос о функциональной роли повторов остался за пределами нашего исследования, хотя для OCT4 ранее уже озвучивалось предположение о существовании альтернативных сайтов связывания, в том числе имеющих структуру повторов или палиндромов [Tantin и др., 2008].

Мы подсчитали AUC ROC по объединенному контрольному набору пиков для всех моно- и динуклеотидных мотивов, выявленных *de novo*, и выяснили, что

результаты слабо отличаются, но мотивы на основе данных [Chen и др., 2008] являются наилучшими для OCT4 и SOX2, и вторыми по успешности для NANOG, с минимальными отличиями в значениях AUC между лидерами. Эти репрезентативные мотивы далее в ходе сравнительных тестов называются ChIPMunk и diChIPMunk, чтобы отличать их от оригинальных, представленных в работе [Chen и др., 2008].

Результаты сравнения мотивов с помощью ROC-кривых представлены на правых панелях на **Рисунке 34**. Значения AUC ROC приведены в легенде.

5.2.1.5. Тройственный композитный элемент OCT4-SOX2/NANOG

По сути, все *de novo* мотивы представляют собой варианты композитного мотива OCT4-SOX2, который наилучшим образом отражает совместное связывание OCT4 и SOX2, характерное для плюрипотентных клеток [Mistri и др., 2015]. Интересно, что мотив для NANOG практически повторяет мотив OCT4-SOX2. В принципе, идентификация композитного элемента OCT4-SOX2 в ChIP-Seq NANOG может быть своего рода артефактом, вызванным прямым перекрытием регионов связывания регуляторов и множеств их генов-мишеней, что подтверждается данными ChIP-Seq. Более того, использование «контрастных» методов идентификации мотивов –все-таки позволяет выявить собственный мотив NANOG (похожий на известные и согласующийся с мотивами других гомеодоменных белков), если в качестве контроля использовать данные по связыванию OCT4 (но не SOX2) [Maaskola, Rajewsky, 2014]. Тем не менее, согласно полученным нами ROC-кривым, среди всех мотивов только мотивам, похожим на OCT4-SOX2, удается успешно распознать сайты связывания *in vivo* (за исключением, в ограниченной степени, мотива HOMER). По всей видимости, в регуляторных районах, связываемых NANOG, присутствует какое-то число его «истинных» собственных сайтов, но большинство регионов связывания содержат именно композитный элемент. Мы сформулировали идею тройственного OSN-композитного элемента OCT4-SOX2/NANOG (нижняя панель на **Рисунке 34**), в котором NANOG связывает участок, перекрывающийся с боксом SOX2, и, вероятно, препятствует устойчивому связыванию гетеродимера OCT4-SOX2. Это согласуется с антагонизмом OCT4 и NANOG в модели генной сети, и подтверждается независимыми исследованиями совместного участия OCT4 и NANOG в регуляции экспрессии [Bin Le и др., 2014].

Мы провели аннотацию композитных элементов в нескольких локусах ключевых генов плюрипотентности и наложили как пики ChIP-Seq, так и данные по эволюционной консервативности. Действительно, предсказанные в пиках сайты связывания в составе композитных элементов в ряде случаев колокализуются с эволюционно-консервативными районами, см. **Рисунок 35**. Вопрос о систематической локализации отдельных сайтов связывания NANOG в составе и вне состава композитных элементов пока остается открытым.

В заключение хочется отметить два замечательных момента.

Во-первых, факт перекрытия сайтов SOX2 и NANOG в композитном элементе с OCT4 был описан ранее для дистального энхансера OCT4 в работе [Young, 2011] (см. врезку).

Во-вторых, NANOG и SOX2 способны самостоятельно взаимодействовать без участия OCT4 и связывать собственный характерный композитный элемент, похожий на, но не идентичный OSN-мотиву [Gagliardi и др., 2013]. Таким образом, в реальности мы имеем дело с интерференцией большего числа сигналов, и подробная декомпозиция мотивов в ChIP-Seq для ключевых факторов плюрипотентности все еще представляет интерес.

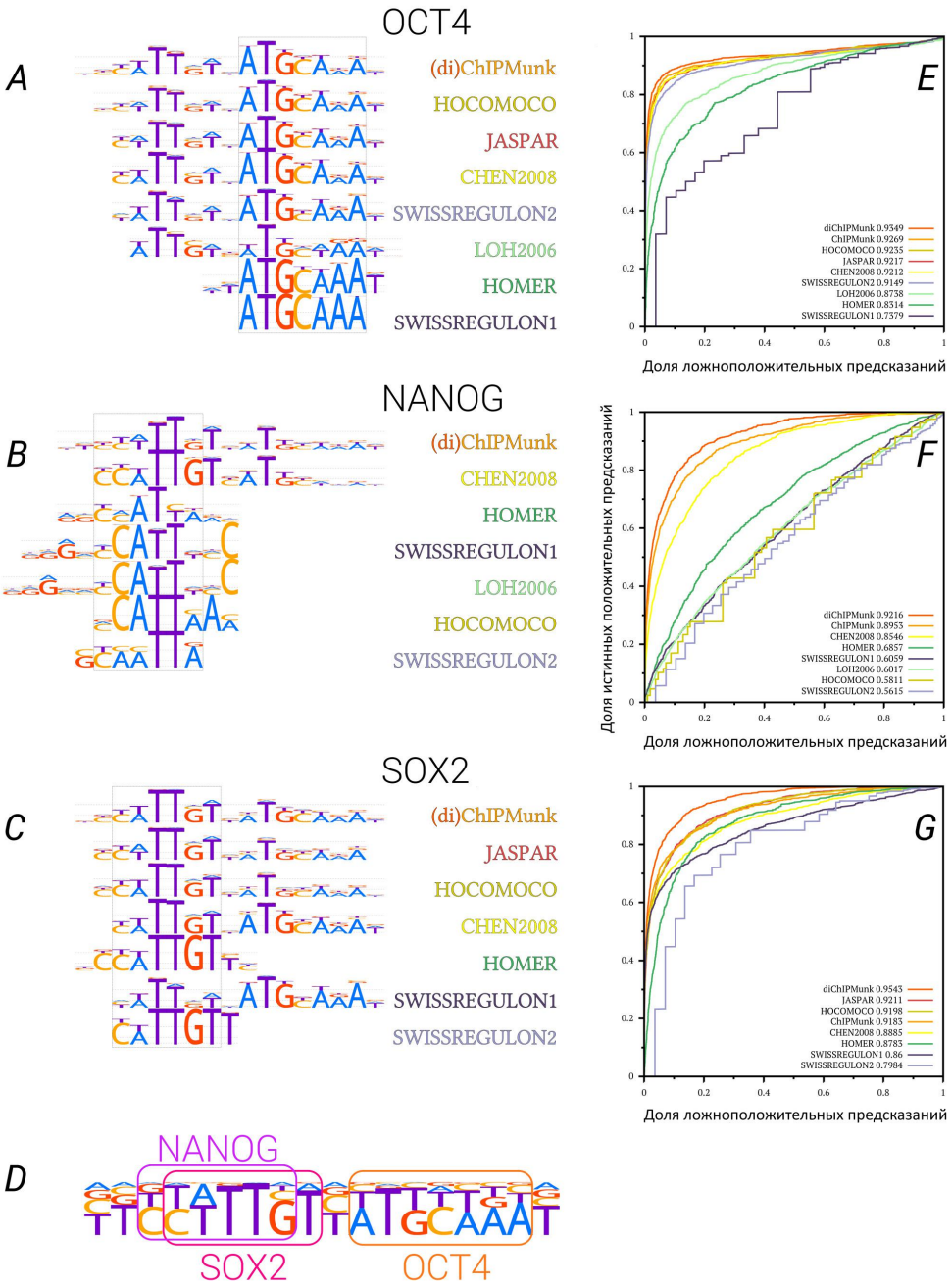


Рисунок 34. Модели мотивов факторов плюрипотентности.

(*левые панели*) Известные и новые модели мотивов для OCT4 (A), NANOG (C) и SOX2 (D). В большинстве случаев фланки мотива соответствуют консенсусу связывания кофактора, а мотив – композитному элементу OCT4/SOX2. Интересно, что это верно и для мотивов NANOG, основанных на ChIP-Seq данных. (*правые панели*) ROC-кривые для сравнения качества мотивов. Мотивы NANOG без OCT4-бокса плохо распознают сайты в пиках ChIP-Seq. Для оценки истинных положительных предсказаний использован объединенный контрольный набор данных для мыши. (*нижняя панель* E) Предлагаемый консенсус тройственного композитного элемента OCT4-SOX2/NANOG. Рисунок адаптирован из работы [Papatsenko и др., 2015].

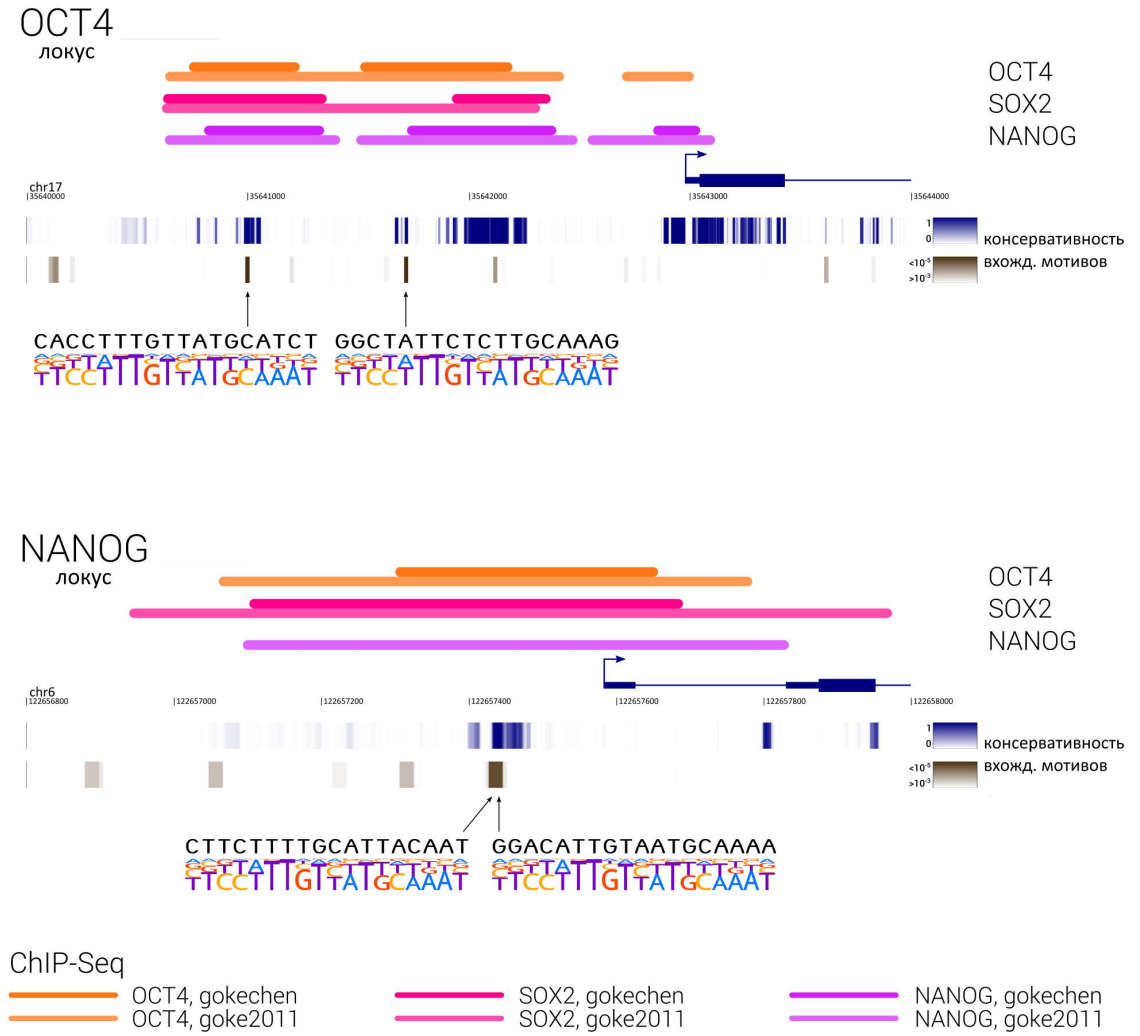


Рисунок 35. Локализация участков связывания OCT4, SOX2 и NANOG на карте локусов OCT4 (*верхняя панель*) и NANOG (*нижняя панель*) в геноме мыши.

Показаны ChIP-Seq пики из различных источников и аннотация композитных элементов (шкала по *P*-значениям) с наложенным треком эволюционной консервативности. Рисунок построен по данным, проанализированным в работе [Papatsenko и др., 2015].

5.2.2. Использование независимых экспериментальных данных для оценки точности представления мотивов сайтов связывания

Обилие данных высокопроизводительных методов, в том числе ChIP-Seq, позволяет проводить масштабный анализ сайтов связывания в котором, однако, всякий раз лучше представлены наиболее сильные и достоверные сайты связывания, не обязательно локализованные в функциональных регуляторных областях. В то же время сайты связывания непосредственно в промоторах и энхансерах не обязаны быть наилучшими в смысле аффинности и, тем более, в смысле сходства с определенной в масштабах полного генома консенсусной последовательностью. В литературе и базах данных присутствуют последовательности сайтов связывания, определенные традиционными «догеномными» методами непосредственно в промоторах или энхансерах генов. Можно ли извлечь из ограниченного набора таких сайтов полезную информацию, а именно, можно ли построить модель, которая будет помогать в повсеместном распознавании сайтов в ChIP-Seq пиках? Можно ли при помощи ограниченных данных выбрать численные пороги распознавания так, чтобы достоверно отделить «настоящие» сайты, узнаваемые в геноме изучаемым белком, от непрямого связывания через кофакторы?

К ответам на эти вопросы мы смогли приблизиться благодаря коллегам из Института цитологии и генетики (ИЦиГ, Новосибирск), которые предложили участвовать в сравнении различных моделей сайтов связывания на основе независимой экспериментальной верификации сайтов методом замедления ДНК-белковых комплексов в гене (EMSA). Наш вклад в работу: построение моделей на основе ChIP-Seq данных и участие в проведении сравнительного тестирования. Альтернативные модели и экспериментальные данные по верификации предсказаний были получены в ИЦиГ [Levitsky и др., 2014].

5.2.2.1. Фактор транскрипции FoxA2 и использованные ChIP-Seq данные

FoxA2 – фактор семейства Fox (forkhead-box, входит в суперкласс ДНК-связывающих доменов типа спираль-поворот-спираль), участвует в регуляции экспрессии генов на различных стадиях жизненного цикла млекопитающих, включая раннее развитие, органогенез и метаболизм у взрослых организмов [Friedman, Kaestner, 2006]. Считается, что FoxA2 является транскрипционным фактором-пионером, и самостоятельно связывает ДНК [Kaestner, 2010]. Можно ожидать, что

большая часть детектируемых ChIP-Seq регионов связывания соответствует прямому связыванию, не опосредованному другими факторами транскрипции.

В качестве основного набора для обучения моделей использовались 4455 пиков ChIP-Seq с покрытием не менее 15 прочтений, определенных в печени взрослой мыши в работе [Wederell и др., 2008]. В качестве дополнительного контроля использовали 4376 регионов связывания с покрытием не менее 10 прочтений по данным ChIP-Seq в клеточной линии HepG2 [Wallerman и др., 2009].

5.2.2.2. Модели сайтов связывания

Для анализа сайтов связывания использовали четыре подхода. Две модели были построены с помощью ChIPMunk и diChIPMunk на полном основном наборе ChIP-Seq пиков и еще две модели были обученные по догеномным данным.

Выравнивания ChIPMunk и diChIPMunk отдельно оптимизировали по длине [Levitsky и др., 2007], таким образом, итоговые модели (20 и 28 п.н.) включали не только канонический консенсус (TRTTTRYH в IUPAC нотации), но и протяженные фланкирующие районы. Альтернативные модели были построены коллегами из ИЦиГ на курированной выборке из 53 сайтов связывания белков семейства FoxA (по данным ДНКазного футпринтинга либо EMSA). Выравнивание последовательностей по консенсусу использовали для построения весовой матрицы (оПВМ, oPWM) и модели SiteGA (генетический алгоритм для выявления зависимостей в сайтах связывания, в том числе между удаленными нуклеотидами). Оптимальная длина выравнивания для oPWM была определена тем же методом, что и для ChIPMunk-моделей (кросс-валидация методом «складного ножа»-jackknife).

5.2.2.3. Тестирование и результаты

Первый технический вопрос, на который мы хотели получить ответ: насколько хорошо справляются ChIP-Seq и продвинутые не-ChIP-Seq модели с распознаванием сайтов связывания в независимом ChIP-Seq контроле (в нашем случае это данные связывания в HepG2 клетках). Для этого мы использовали ROC-кривые (**Рисунок 36**), которые демонстрируют значительное превосходство diChIPMunk в распознавании сайтов в ChIP-Seq пиках. Модели, построенные на небольшой курированной выборке сайтов, по качеству распознавания сайтов сравнимы с моделями TRANSFAC; модели JASPAR, основанные на ChIP-Seq данных,

неотличимы от результатов ChIPMunk. Полученные результаты хорошо согласуются с ожидаемыми.

Более глубокий анализ требует выделения сайтов связывания, уникально распознаваемых каждой конкретной моделью из четырех. Мы выбрали 466 предсказанных сайтов связывания (наилучшие предсказания каждой модели) в пиках, перекрывающихся с промоторами генов (1000 п.н. «апстрим» – в 5’ областях относительно стартов транскрипции). Для этих сайтов были построены диаграммы рассеивания для оценок различных пар моделей. Удалось обнаружить массу примеров, когда оценки различных методов были несогласованы. 64 сайта с несогласованными оценками моделей были выбраны для верификации с помощью EMSA. Использование канонического сайта связывания FoxA2 из промотора гена транстиренина [Lai и др., 1991] (TTR) в качестве положительного контроля и олигонуклеотида с сайтом PPAR в качестве отрицательного контроля позволило получить нормированную количественную оценку аффинности (где 1 соответствует связыванию TTR и 0 соответствует неспецифическому связыванию PPAR), которая затем использовалась для выбора пороговых значений оценок.

На диаграммах рассеивания комбинация мотивов, построенных SiteGA и diChIPMunk, наилучшим образом отделила слабые сайты и не-сайты от наиболее достоверных [Levitsky и др., 2014]. В частности, это позволяет утверждать, что знание удаленных зависимостей даже в ограниченном «учебном» наборе сайтов (соответствует модели SiteGA) действительно позволяют точнее описать связывание белка *in vivo*. Чтобы подкрепить эту мысль был проведен анализ доли распознаваемых пиков контрольного ChIP-Seq при фиксированных порогах распознавания, которые для всех моделей были установлены на уровне слабых сайтов по данным EMSA (с относительной EMSA-оценкой 0.25). И действительно, на таком пороге мотив diChIPMunk самостоятельно выделял сайты в большей части пиков, но в то же время, добавление SiteGA позволяло успешно распознать сайты еще в 5-10% выборки, см. **Рисунок 37**.

Резюмируя: диПВМ diChIPMunk хорошо распознает основную долю сайтов, но ряд регионов связывания «ускользает» от выравниваний, и связывание FoxA2 с ними может быть достоверно объяснено моделью с учетом удаленных зависимостей, причем, даже при обучении по ограниченной «догеномной» выборке сайтов.

Остается открытым вопрос, удастся ли для построения таких моделей обойтись данными ChIP-Seq [Keilwagen, Grau, 2015; Siebert, Söding, 2016] или включение альтернативных источников данных даже ограниченного объема будет стабильно приносить пользу, например, за счет учета слабых-альтернативных сайтов.

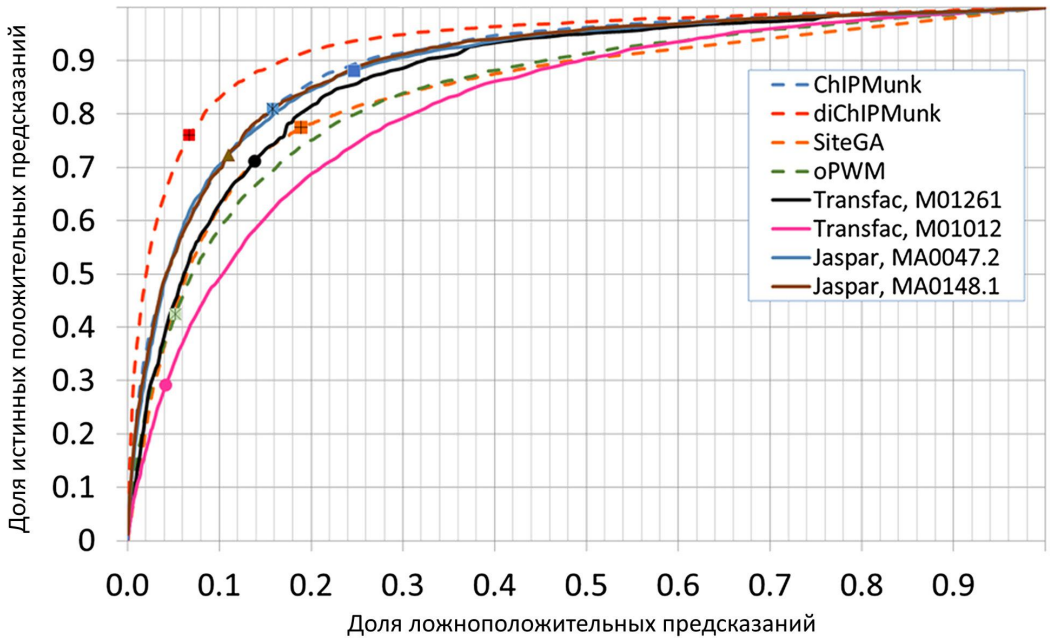


Рисунок 36. ROC-кривые для различных моделей мотивов FoxA2.

Данные ChIP-Seq в клетках HepG2 использованы в качестве положительной контрольной выборки. В качестве отрицательного контроля взяты последовательности пиков с перемешанными нуклеотидами. В сравнение включены матрицы из TRANSFAC и JASPAR. Маркеры на графиках соответствуют порогам оценок, определенным по сайтам, верифицированным EMSA. Рисунок адаптирован из работы [Levitsky и др., 2014].

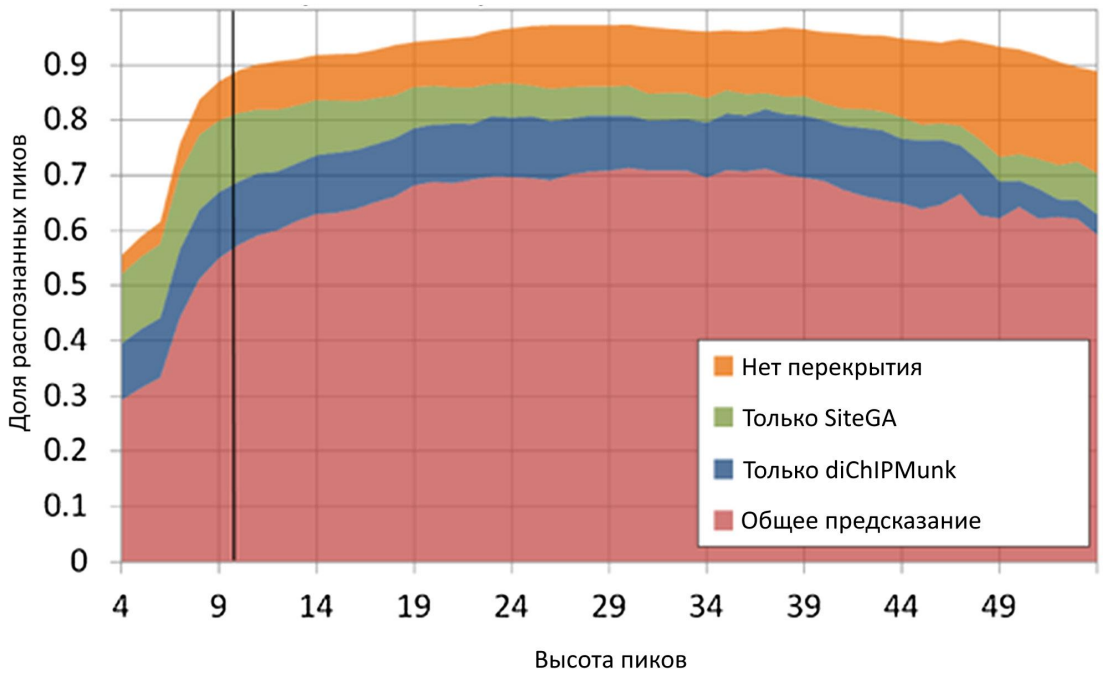


Рисунок 37. Доля пиков, содержащих надпороговые вхождения мотивов FoxA2. Пороги распознавания мотивов SiteGA и diChIPMunk определены по результатам EMSA. По оси Y отложена доля пиков с распознанными сайтами среди всех пиков высотой не ниже значения по оси X. Раздельно показаны доли пиков с надпороговыми предсказаниями для обеих моделей (перекрывающихся и не перекрывающихся в конкретной последовательности) или только одной из двух моделей. Для каждой модели всякий раз взято только наилучшее вхождение мотива.

5.2.3. Кластеризация сайтов связывания фактора транскрипции Spi1 и регуляция экспрессии генов при эритролейкемии

В этом разделе обсуждается эффект кластеризации сайтов связывания фактора транскрипции Spi1 на экспрессии близлежащих генов-мишеней.

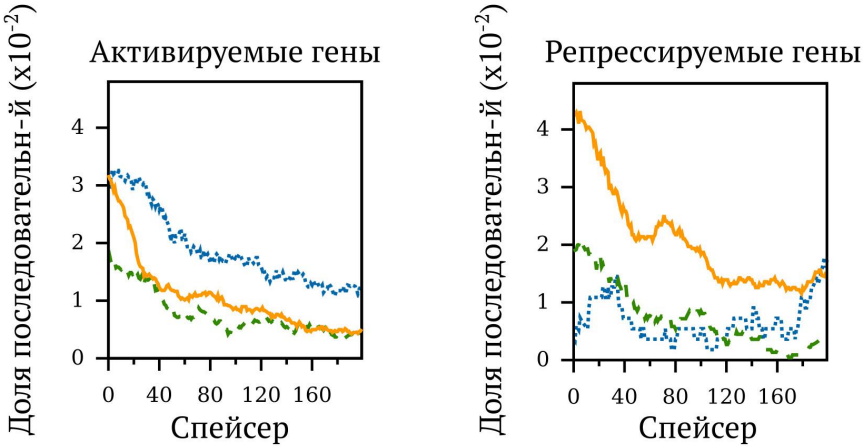
Фактор транскрипции Spi1/PU.1 является одним из ключевых регуляторов экспрессии генов в клетках костного мозга и B-лимцофитах [Iwasaki и др., 2005]. Оверэкспрессия Spi1 блокирует дифференцировку предшественников эритроцитов и приводит к эритролейкемии [Moreau-Gachelin и др., 1996]. Spi1 относится к ETS-семейству факторов транскрипции (суперкласс спираль-поворот-спираль) и связывает характерный консенсус с коровым элементом GGAA. Чтобы понять, как именно нарушается работа генной сети при оверэкспрессии Spi1, наши коллеги из института Кюри (Париж) провели серию опытов на клетках селезенки трансгенной мыши, оверэкспрессирующих Spi1 [Ridinger-Saison и др., 2012]. С помощью комбинации ChIP-Seq и экспрессионного профилирования на микрочипах удалось не только установить полногеномный профиль сайтов связывания, но и соотнести связывание фактора транскрипции с изменением экспрессии близлежащих генов.

Наш вклад в эту работу – построение модели мотива связывания Spi1 и анализ характерного взаимного расположения сайтов связывания в регуляторных областях. Мотив связывания Spi1 был построен на основе полного набора пиков ChIP-Seq (17781 регионов определенных с помощью MICSA [Boeva и др., 2010] и FindPeaks [Fejes и др., 2008]). Полученный мотив хорошо согласуется с литературными данными, по оценке ChIPMunk вхождения мотива присутствовали почти в 90% пиков. Затем пики были разбиты на группы (а) в соответствии с геномной локализацией: промоторы, включая ближайшие 3’-окрестности стартов (-1.5 тыс. п.н. до +2 тыс. п.н. вокруг старта инициации), потенциальные энхансеры (вплоть до -30 тыс. п.н. до старта), внутригенные и межгенные участки; и (б) в соответствии с предполагаемой функцией (активация либо ингибирование экспрессии ближайшего гена). Фиксированный порог на изменение экспрессии в полтора раза выявил 672 гена, активируемых или ингибируемых при нокдауне Spi1. 70-80% генов содержали пики Spi1 в ближайшей окрестности (в регионе от 30 тыс. п.н. в 5’ область до 5 тыс. п.н. в 3’ область).

Для поиска вхождений мотива использовали порог, соответствующий *P*-значению в 0.0001(6), что соответствует случайным предсказаниям примерно в 10% двуцепочечных последовательностей длины 300 п.н., близкой к характерной длине пиков ChIP-Seq.

Задача состояла в определении структуры регуляторной последовательности, определяющей результат связывания Spi1: активация или ингибирование экспрессии генов-мишеней. Мы изучили позиционные предпочтения Spi1 в пиках ChIP-Seq и обнаружили замечательный факт: «знак» эффекта, оказываемого Spi1, зависит от взаимной ориентации парных сайтов связывания. Связывание одинаково ориентированных парных сайтов в промоторах чаще ведет к активации экспрессии, но только для промоторов, не пересекающихся с CpG-островками (это может быть связано с типом промоторов или связыванием кофакторов). Обратно, связывание Spi1 тандемных сайтов в энхансерных областях приводит к ингибированию экспрессии, см. **Рисунок 38**. Нельзя полностью исключить, что парные сайты в каких-то случаях соответствуют посадке и других членов ETS-семейства помимо Spi1, но нам представляется примечательным сам факт различной функциональной роли гомотипических пар сайтов с одинаковой ориентацией в зависимости от геномной локализации. Особенно интересно, что пары сайтов с противоположной ориентацией редко встречаются во всех типах геномных областей.

Пары сайтов связывания Spi1-Spi1 в идентичной (прямой) взаимной ориентации



Пары сайтов связывания Spi1-Spi1 в обратно-комплементарной ориентации

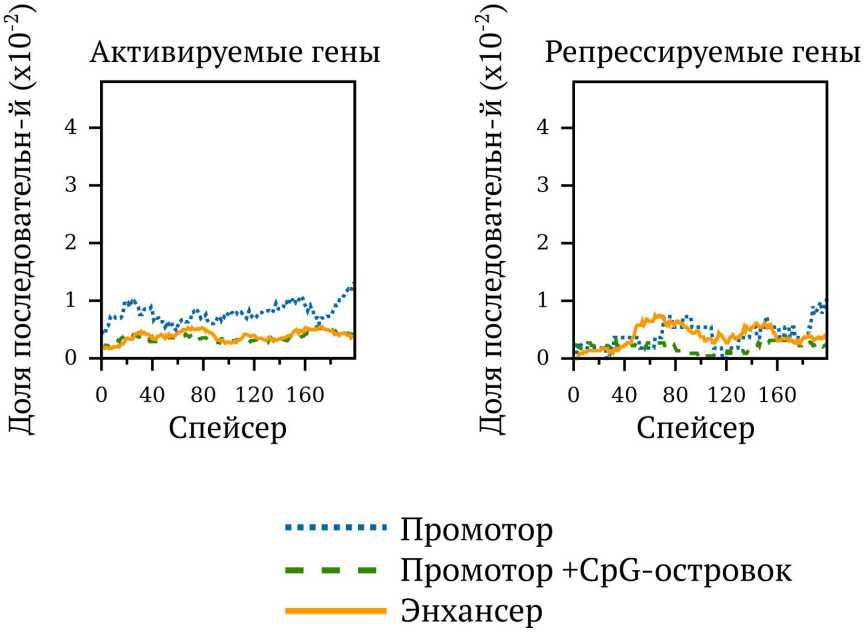


Рисунок 38. Предпочтительные расстояния между парами сайтов Spi1 в пиках ChIP-Seq в тандеме (*верхняя панель*) и в обратно-комплементарной ориентации (*нижняя панель*).

Функциональные категории пиков: (*сплошная линия*) потенциальные дистальные энхансеры; (*пунктирная линия*) промоторы с CpG-островком; (*линия с точками*) прочие промоторы. Ось X: расстояние между вхождениями мотива Spi1. Ось Y: доля пиков ChIP-Seq, содержащих пару сайтов Spi1 на заданном расстоянии. Данные взяты из работы [Ridinger-Saison и др., 2012]. Рисунок адаптирован из обзора [Kulakovskiy, Makeev, 2013].

5.2.4. Взаимосвязь транскрипции и трансляции мРНК-мишеней сигнального каскада mTOR

В этом разделе представлен анализ терминального олигопиримидинового мотива мРНК, колокализованного с сайтами инициации транскрипции генов, регулируемых сигнальным каскадом mTOR на уровне трансляции.

5.2.4.1. Терминальный олигопиримидиновый мотив и регуляция трансляции в ответе на сигнальный каскад mTOR

Киназа mTOR (mammalian target of rapamycin, мишень рапамицина у млекопитающих) является одним из ключевых регуляторов клеточного роста и пролиферации у высших эукариот [Johnson, Rabinovitch, Kaeberlein, 2013; Wang, Proud, 2011]. Каскад mTOR играет важную роль в онкогенезе [Topisirovic, Sonenberg, 2011; Zoncu, Efeyan, Sabatini, 2011], что стимулирует исследования молекулярных механизмов, в том числе, роли mTOR в регуляции трансляции. Достаточно давно изучались особенности 5’ НТО (нетранслируемых областей) мРНК, трансляция которых зависит от клеточного роста [Meyuhas, 2000]. Более двадцати лет назад в некоторых мРНК был обнаружен короткий 5’-Терминальный ОлигоПиримидиновый мотив (ТОП, TOP, terminal oligopyrimidine tract) [Avni, Biberman, Meyuhas, 1997; Hornstein и др., 1999; Jefferies, Thomas, 1994; Terada, 1994]. Сегодня известно, что эволюционно консервативный ТОП [Perry, 2005] содержат многие мРНК рибосомных белков и факторов трансляции, а мотив действительно представляет собой олигопиримидиновую последовательность длины 5-14 нуклеотидов, локализованную непосредственно на 5’ конце 5’ НТО [Meyuhas, 2000].

Параллельно был обнаружен специальный класс промоторов, зависимых от TCT-мотива, строго локализованного в позиции, где непосредственно стартует инициация транскрипции [Parry и др., 2010; Rach и др., 2011]. Транскрипция мРНК рибосомных белков с TCT-промоторов позволяет говорить о пиримидиновом мотиве «двойного назначения», участвующем в регуляции и транскрипции, и трансляции. Анализ ТОП-мотива несколько осложнен тем фактом, что точность аннотации сайтов инициации транскрипции в стандартных базах данных долгое время была ограничена, что не позволяло точно определить 5’ конец мРНК для большого набора генов [Yamashita и др., 2008].

Данные рибосомного профайлинга (Ribo-Seq) по изменению эффективности трансляции при ингибировании mTOR повторно подняли вопрос о повсеместном наличии и функциональной необходимости терминального ТОП мотива для регуляции трансляции. На основании анализа сотен генов-мишеней mTOR было предположено существование нетерминальных и вырожденных мотивов ТОП [Thoreen и др., 2012] либо внутренних олигопиримидиновых трактов (pyrimidine-rich translational elements, PRTE) в удалении от 5’ конца 5’ НТО [Hsieh и др., 2012].

Полногеномные данные по экспериментальному картированию сайтов связывания, например, представленные в базе данных DBTSS[47] [Yamashita и др., 2012], не до конца проясняли картину в силу неполноты информации по клеточным типам и ограниченной согласованности с геномной аннотацией.

Новый источник количественных данных по активности стартов транскрипции (TSS, transcription start site) появился благодаря технологии HeliScopeCAGE (Cap Analysis of Gene Expression using Helicos single-molecule sequencing). Это кэп-анализ экспрессии генов с использованием технологии Helicos для секвенирования одной молекулы, без использования амплификации ПЦР [Kanamori-Katayama и др., 2011]. HeliScopeCAGE позволяет проводить полногеномный количественный анализ стартов транскрипции с разрешением вплоть до отдельных нуклеотидов. Естественным кажется желание объединить точные данные по трансляции (Ribo-Seq) и транскрипции (HeliScopeCAGE), чтобы прояснить вопрос о мотивах, участвующих в регуляции mTOR-ответа на уровнях транскрипции и трансляции. В нашей работе [Eliseeva и др., 2013] мы применили методы анализа мотивов, чтобы заново идентифицировать мотив ТОП и надежно оценить его локализацию относительно 5’ концов 5’ НТО и стартов транскрипции.

5.2.4.2. ТОП-мотив, идентифицированный *de novo*, хорошо согласуется с известным

С помощью ChIPMunk в 5’ НТО мРНК-мишеней mTOR мы определили основной CU-богатый мотив. Для этого использовали 250 транскриптов 142 генов-трансляционных мишеней mTOR человека из работы [Hsieh и др., 2012]. Мотив идентифицировали в 5’ НТО, расширенных на 100 п.н. в 5’ область относительно

[47] DataBase of Transcriptional Start Sites. http://dbtss.hgc.jp/

геномной аннотации UCSC, чтобы учесть возможные ошибки аннотации стартов. Оптимальный мотив длины 14 хорошо согласуется с известным ранее представлением ТОП сравнимой длины с мажорным цитозином в UCU(TCT) контексте [Perry, 2005; Yamashita и др., 2008]. Вхождения мотива были идентифицированы ChIPMunk в 214 из 250 5’ НТО (для 126 из 142 генов). Стоит отметить, что эти вхождения часто содержали как минимум одно пуриновое основание (для 184/108 5’ НТО/генов), а во многих случаях (для 84/58 5’ НТО/генов) пуриновое основание было локализовано в 8 средних позициях мотива, что подчеркивает, что жесткая безразрывная пиримидиновая структура не является обязательной. По аналогии с ТОП-мотивом в мРНК мы называем соответствующий участок ДНК ОП-мотивом (поскольку в ДНК он не является терминальным).

5.2.4.3. ОП/ТОП-мотив обладает выраженными позиционными предпочтениями

Стандартная генная аннотация не позволяет корректно изучать локализацию ОП/ТОП

Для оценки позиционных предпочтений ТОП в 5’ НТО необходима точная аннотация стартов транскрипции. Значительным шагом вперед является использование прямых экспериментальных данных, например представленных в DBTSS; либо курированной переаннотации генов и транскриптов GENCODE [Harrow и др., 2012]. В этом исследовании мы пользовались данными HeliScopeCAGE с однонуклеотидным разрешением [Kanamori-Katayama и др., 2011]. Стандартная аннотация стартов транскрипции (например, по данным, UCSC Genome Browser [Fujita и др., 2011]) для mTOR-мишеней является более точной – в сравнении с полным набором белок-кодирующих генов, см. **Рисунок 39**. Однако, даже для mTOR-мишеней точность невысока: менее чем для половины генов аннотированный старт находится в пределах 10 п.н. от максимума сигнала CAGE (который соответствует старту мажорной изоформы мРНК). То есть, стандартная аннотация стартов транскрипции делает детализированный анализ локализации мотивов невозможным.

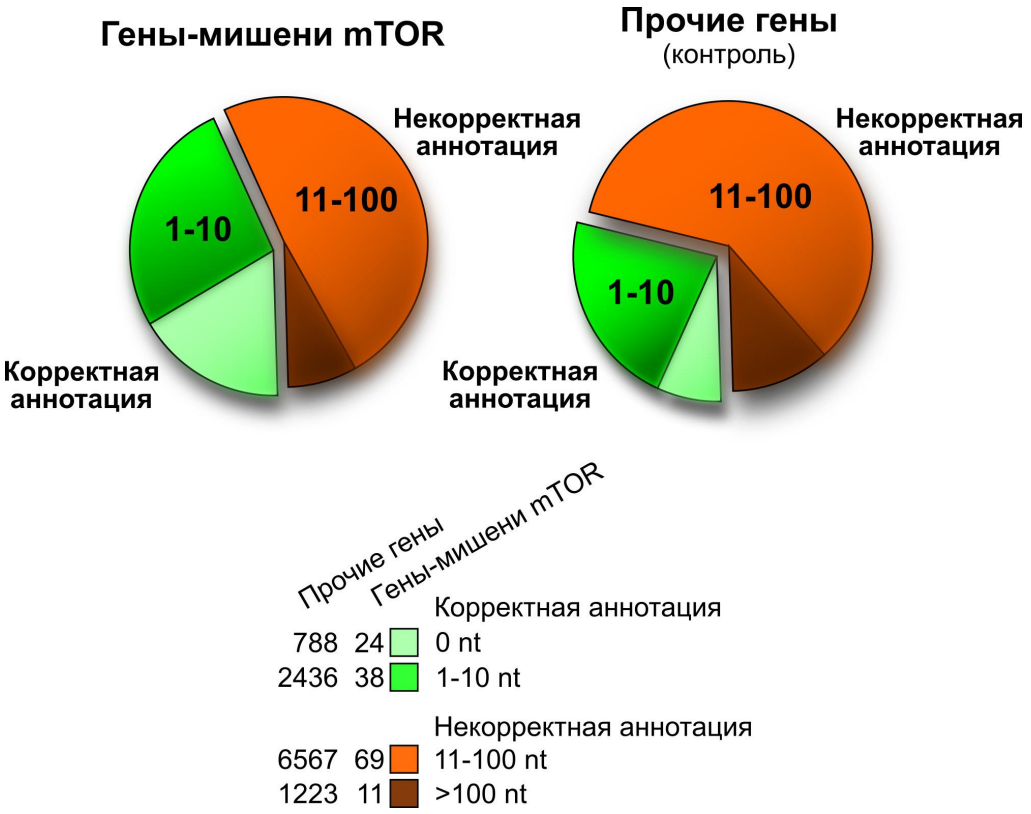


Рисунок 39. Число генов, для которых аннотированный и верифицированный старты транскрипции находятся на заданном расстоянии. Аннотация UCSC hg18 сравнивается с максимумами профилей HeliScopeCAGE. Для каждого транскрипта рассматривается окрестность от 100 п.н. в 5’ область относительно аннотированного старта до 3’ конца 5’ НТО. Для каждого гена выбрано наименьшее (наилучшее) расстояние среди всех аннотированных транскриптов. Рисунок адаптирован из работы [Eliseeva и др., 2013].

Реальная локализация ОП/ТОП-мотивов

Чтобы выяснить реальную локализацию олигопиримидиновых мотивов мы использовали данные CAGE: подсчитывали число 5’ НТО с вхождения ОП-мотива в конкретной позиции относительно максимума CAGE-сигнала (наибольшее число картированных Helicos-чтений, где каждое чтение соответствует 5’ концу одной молекулы мРНК, извлеченной непосредственно за кэп). Далее мы оценили статистическую значимость ассоциации между принадлежностью транскрипта множеству mTOR-мишеней и наличию вхождений ОП-мотива на конкретной позиции 5’ НТО (в качестве нулевой гипотезы использовали предположение, что частота и положение ОП-мотива не отличаются для mTOR-мишеней и прочих транскриптов, которые выполняли роль нейтрального контроля).

Обогащение ОП в 5’ области и с перекрытием старта имеет высокую статистическую значимость (*P*-значение точного теста Фишера <<0.05) вне зависимости от порога оценок ОП-мотива. Обогащение ОП-трактов в 3’ области относительно старта (т.е. внутри 5’ НТО) имело меньшую значимость, но в позициях от +1 до +4 все еще показывало значимые $P < 0.05$ (в зависимости от порога). Пример позиционного распределения показан на **Рисунке 40**.

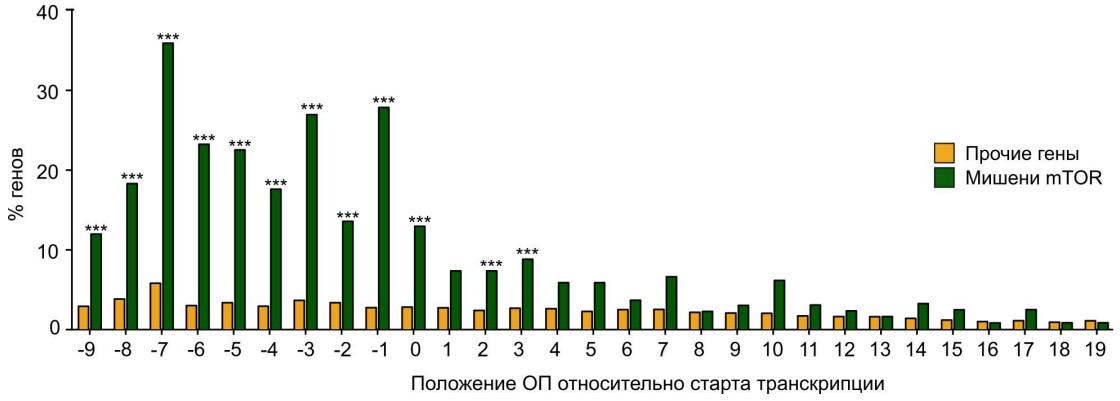


Рисунок 40. Доля генов с вхождениями олигопиримидинового мотива на различных расстояниях от максимума CAGE-профиля.

Вхождения мотива определены на уровне *P*-значений 0.005. Значимые отличия промаркированы *** (точный тест Фишера для 5' НТО достаточной длины, $P < 0.05$ после поправки Холма на множественное тестирование). Ноль по оси X соответствует положению максимального CAGE-сигнала. Рисунок адаптирован из работы [Eliseeva и др., 2013].

Наконец, мы изучили общие позиционные предпочтения ОП-мотива, подсчитав ОП-вхождения не далее 19 п.н. от верифицированных стартов (в «головах» 5’ НТО) и, отдельно, в остающихся «хвостах». По сравнению с контрольными данными, вхождения ОП-мотива были существенно перепредставлены в «головах» ($P << 0.05$) но не в «хвостах» 5’ НТО mTOR-мишеней ($P >> 0.05$). Попытки *de novo* идентификации дополнительных пиримидин-богатых мотивов в хвостах 5’ НТО не увенчались успехом. Таким образом, нам не удалось найти аргументов ни в пользу существования «похожих-на-ТОП» (TOP-like) мотивов [Thoreen и др., 2012] ни в пользу обогащения пиримидин-богатыми трактами «глубин» 5’ НТО [Hsieh и др., 2012].

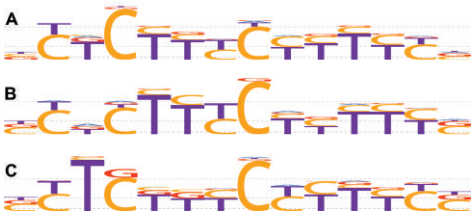
Олигопиримидиновые мотивы отличаются для узких и широких стартов

Гены рибосомных белков и трансляционного аппарата транскрибируются с TCT-промоторов, особого класса, характеризуемого TCT-консенсусом и окружающим его олигопиримидиновым мотивом. Точная инициация транскрипции с TCT-промоторов [Rach и др., 2011] гарантирует наличие ТОП-мотива непосредственно на 5’ конце мРНК.

В то же время, многие mTOR-мишени транскрибируются с широких промоторов. Используя количественные данные HeliScopeCAGE для мишеней mTOR мы оценили типичную протяженность региона, соответствующего 5’ концам основных транскриптов. Для начала мы подсчитали долю мРНК, транскрибируемых с конкретной позиции генома, для которой CAGE-сигнал максимален в локальной окрестности аннотированного старта. Выяснилось, что для более чем половины mTOR-мишеней с одной «мажорной» позиции транскрибируется менее половины пула мРНК. Таким образом, mTOR-мишени, даже с учетом TCT-промоторов, плохо соответствуют упрощенному представлению старта транскрипции как точечного объекта.

Приняв это во внимание, мы примерно оценили ширину региона инициации транскрипции («ширину старта») для конкретного гена как минимальную протяженность геномного региона, покрывающего старты транскрипции для как минимум 2/3 пула конкретной изоформы мРНК, см. **Рисунок 41**. Затем мы разделили все транскрипты на два класса: 167 последовательностей расширенных 5’ НТО, соответствующих транскрипции с узких стартов (10 и менее п.н.); и 83

последовательности, транскрибируемые с широких стартов (более 10 п.н.). Результатом идентификации мотивов *de novo* стали олигопиримидиновые мотивы с вхождениями в 138/63 последовательностей 5’ НТО, соответствующих узким/широким стартам, см. врезку (**A** – усредненный мотив, **B** – узкие старты, **C** – широкие старты). Все три мотива похожи, за исключением мажорного T для варианта, извлеченного из 5’ НТО для широких стартов. Интересно, что этот мотив похож на «внутренней» пиримидин-богатый мотив (PRTE); есть прямые основания утверждать, что идентификация мотива PRTE «внутри» 5’ НТО стала продуктом неточной аннотации широких стартов инициации транскрипции.



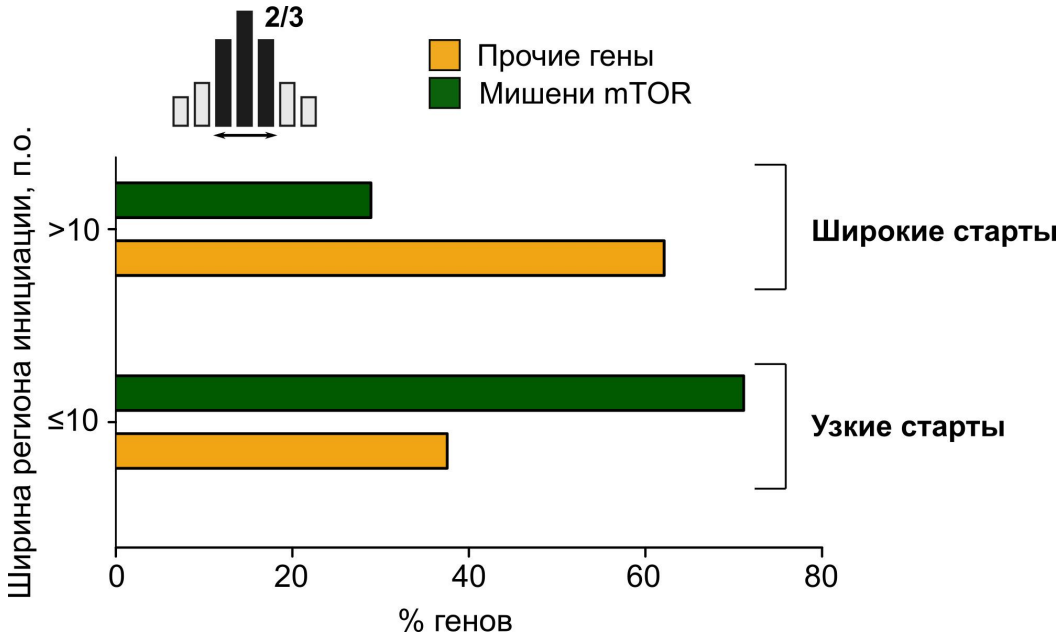


Рисунок 41. Процент генов (ось X), имеющих старт (область инициации транскрипции) заданной ширины (ось Y).

Ширина старта определена как минимальная протяженность региона, в котором находится как минимум 2/3 от суммарного CAGE-сигнала. Рисунок адаптирован из работы [Eliseeva и др., 2013].

Широкие регионы инициации транскрипции конкретных mTOR-мишеней могут продуцировать ТОП и не-ТОП мРНК

ОП-мотив окружает старт и точная транскрипция с ТСТ-промоторов гарантирует наличие ТОП-последовательности на 5’ конце мРНК. Форма профиля CAGE (т.е. относительная активность инициации транскрипции с конкретных позиций в рамках одного региона инициации) определяет состав конкретных вариантов 5’ НТО в пуле мРНК. Для широких стартов ОП-мотив может покрывать только часть региона инициации транскрипции. Экстремальный пример – широкие мультимодальные старты с выраженными «модами» среди которых только часть покрыта олигопиримидиновым трактом. Локальное переключение активности стартов на транскрипционном уровне может влиять на присутствие ТОП в мРНК, то есть регуляция транскрипции может определять дальнейшую регуляцию трансляции. Примеры CAGE-профиля для избранных мишеней mTOR показаны на **Рисунке 42**.

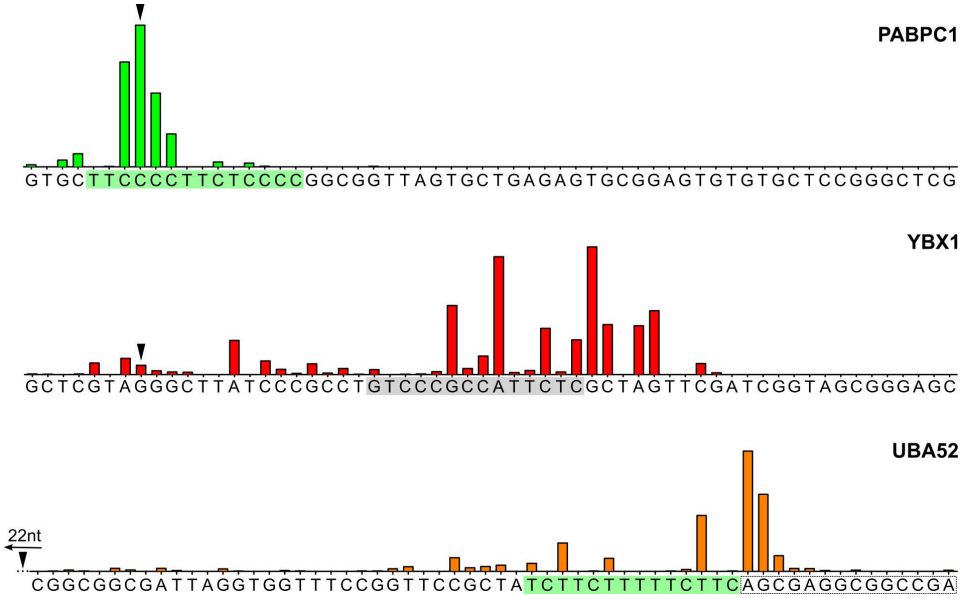


Рисунок 42. CAGE-сигнал и вхождения ОП-мотива в расширенные 5’ НТО трех мРНК-мишеней mTOR: PABPC1 (*верхняя панель*), YBX1 (*средняя панель*), UBA52 (*нижняя панель*).

Черным треугольником показан старт транскрипции согласно аннотации генома человека (UCSC hg18). Наилучшие вхождения ОП-мотива в последовательность выделены цветным фоном. PABPC1 выглядит как классический ТОП-ген, YBX1 иллюстрирует специальный случай с широким стартом и слабо похожим на консенсус олигопиримидиновым трактом, UBA52 потенциально обладает мультимодальным стартом транскрипции. Рисунок адаптирован из работы [Eliseeva и др., 2013].

5.2.4.4. Методические замечания

Подготовка последовательностей 5’ НТО

Набор белок-кодирующих генов, картированных на идентификаторы Entrez (Gene ID) был извлечен согласно HGNC[48] (human gene names consortium, [Gray и др., 2013]). Аннотация соответствующих транскриптов и базовая аннотация стартов были взяты из геномного браузера UCSC согласно референсной сборке генома человека hg18. Последовательности интронов были исключены, а 5’ НТО были расширены на -100 п.н. в 5’ область относительно аннотированного старта.

Список из 144 генов-мишеней mTOR на уровне трансляции, определенных в клетках PC3 с помощью рибосомного профайлинга, был взят из работы [Hsieh и др., 2012]. 142 гена соответствовали двум критериям: имели непустой профиль HeliScopeCAGE (см. ниже) в окрестности 5’ конца 5’ НТО и были однозначно картированы по идентификаторам генов UCSC-Entrez. Этим 142 генам соответствовало 250 последовательностей 5’ НТО (в силу наличия альтернативных аннотаций транскриптов). Белок-кодирующие гены, удовлетворяющие критериям, но не входящие в список мишеней, использовались в качестве контрольного набора данных (17671 5’НТО для 11027 генов).

Подготовка данных HeliScopeCAGE

На момент работы в открытом доступе были доступны данные CAGE для клеточных линий THP-1 и HeLa[49] [Kanamori-Katayama и др., 2011]. Исходный сигнал HeliScopeCAGE пропорционален числу транскрибируемых с конкретной позиции молекул мРНК. Для обоих клеточных линий мы усреднили полногеномные данные по имеющимся репликам, обнулили значения, присутствующие только в одной реплике, и затем усреднили данные по двум клеточным линиям с округлением вниз до ближайшего целого. В настоящее время подробные данные CAGE уже доступны для клеточной линии PC3, и можно отметить, что для большинства классических mTOR-мишеней старты транскрипции имеют близкую форму и локализацию.

Идентификация мотивов de novo

Для идентификации мотивов мы использовали ChIPMunk и, дополнительно, MEME

[48] HGNC database of human gene names. http://www.genenames.org/
[49] http://fantom.gsc.riken.jp/5/suppl/Kanamori-Katayama_et_al_2011/

[Bailey и др., 2015] в качестве дополнительного контроля. ChIPMunk использовался в режиме одноцепочечного поиска с учетом локального динуклеотидного состава, для достижения полной стабильности результатов использовалось в 100 раз больше случайных семян, по сравнению с значениями по умолчанию. Автоматически определенная длина мотива составляла 25 п.н., включая 14 п.н. олигопиримидинового тракта, фланкированного GC-богатыми участками. Фиксированная длина 14 п.н. использовалась для перезапуска ChIPMunk и построения итогового мотива.

Поиск вхождений мотивов в 5' НТО

Для поиска ОП-вхождений фиксированные пороги ПВМ выбирались с помощью MACRO-APE. Три порога (низкий, средний, высокий), соответствовали *P*-значениям 0.005, 0.0005 и 0.00005.

Оценка позиционных предпочтений

Мы подсчитывали число 5' НТО с вхождением ОП/ТОП-мотива в конкретной позиции относительно максимума HeliScopeCAGE пика. Расстояние 0 соответствовало вхождению ОП, начинающемуся строго на позиции максимума (предпочитаемой-мажорной локализации старта). Для каждого гена мы рассматривали все аннотированные транскрипты, ген учитывался как ОП-ген если любой из его транскриптов содержал вхождение ОП в 5' НТО. При оценке обогащения вхождений ОП в конкретных позициях использовалась поправка Холма на множественное тестирование [Holm, 1979].

5.2.4.5. Обсуждение и заключение по разделу

Нам удалось сделать ряд интересных наблюдений о локализации и подтипах олигопиримидиновых мотивов в промоторах узких и широких регионов инициации транскрипции. Тем не менее, использованный подход имеет ряд ограничений. Во-первых, вариативны и ширина старта и длина ОП-мотива, что не соответствует идее безделеционного множественного локального выравнивания и весовой матрице фиксированной длины. Другая проблема – пересечение ТОП-мотива в мРНК с собственной структурой ТСТ-промторов, что осложняет разделение транскрипционного и трансляционного сигнала (в частности, ОП-мотив в обратно-комплементарной форме чрезвычайно похож на сайты связывания белков ETS-

семейства, в т.ч. Spi1, которые часто локализуются непосредственно на старте транскрипции). В-третьих, использованные эксперименты по кэп-анализу экспрессии были проведены для клеточных линий, отличных от PC3, для которой проведен эксперимент по рибосомному профайлингу, т.е. нельзя исключить, что конкретные гены или транскрипты были неправильно отнесены к ОП или не-ОП классу.

Мы выявили десятки mTOR-мишеней, не соответствующих простой ОП-модели: полностью лишенных вхождений олигопиримидинового мотива или имеющих широкие старты транскрипции с размытыми олигопиримидиновыми треками. Например, именно к такому классу относится YBX1 с широким регионом инициации транскрипции и массой вариантов 5’ НТО, один из которых, как предполагалось, обладает функциональным внутренним полипиримидиновым трактом [Hsieh и др., 2012]. Прямой эксперимент [Lyabin, Eliseeva, Ovchinnikov, 2012] показал, что укороченная не-ТОП мРНК YBX1 успешно регулируется mTOR на уровне трансляции, не смотря на отсутствие каких бы то ни было пиримидиновых последовательностей в 5’ НТО. То есть, трансляционный контроль mTOR задействует и другие регуляторные механизмы, не связанные с ТОП-мотивом.

Нам не удалось обнаружить альтернативных мотивов в 3’ области 5’ НТО. Тем не менее, для 5’ НТО mTOR-мишеней по сравнению с контрольным набором в общем характерен пиримидин-богатый нуклеотидный состав и меньшая длина. То есть, вопрос о возможных специальных свойствах или сигналах в последовательностях еще не закрыт. Возможная функциональная роль GC-богатых фланкирующих участков у ОП-мотивов также требует дальнейшего изучения [Biberman, Meyuhas, 1997], как и потенциальная альтернативная регуляция транскриптов с мультимодальных ОП/не-ОП стартов. В частности, зависимая от роста трансляция генов рибосомных белков отличается в различных типах клеток [Avni, Biberman, Meyuhas, 1997]; и в то же время в различных тканях конкретный ген может транскрибироваться с альтернативных стартов порождая ТОП либо не-ТОП-мРНК с различной эффективностью трансляции [Kleene и др., 2003].

Эти вопросы появятся требуют дальнейших детальных исследований, как с привлечением масштабных данных рибосомного профайлинга, так и точечных экспериментов по мутагенезу конкретных вариантов 5’ НТО.

5.2.5. Давление отбора на соматические мутации в сайтах связывания факторов транскрипции в геномах раковых клеток

Соматические мутации в сайтах связывания факторов транскрипции [Jiang и др., 2015; Mathelier и др., 2015b] могут изменять аффинность связывания регуляторов и экспрессию генов-мишеней [Melton и др., 2015], что может приводить к перестройке генных сетей и, потенциально, к злокачественной трансформации клеток. Наиболее известный пример: сайты посадки фактора GABP (член ETS-семейства), возникающие в результате соматических мутаций в промоторе гена теломеразы TERT [Bell и др., 2015], связанные с развитием различных видов рака [Huang и др., 2013; Killela и др., 2013].

Связь с увеличенным риском развития рака уже установлена и для других мутаций в промоторах [Landa и др., 2010; Li и др., 2011b]; можно ожидать, что объем информации о роли некодирующих вариантов в онкогенезе будет продолжать расти вместе с объемом данных по индивидуальной геномике. Например, в недавнем исследовании ассоциации геномных вариантов с риском развития эпителиального рака яичников (epithelial ovarian cancer [Lawrenson и др., 2015]), лишь 2 полиморфизма из почти 300 значимых были локализованы в кодирующих областях, при этом 25 из 300 были найдены непосредственно в сайтах связывания факторов транскрипции. В то же время, для генов-драйверов онкогенеза, найденных в массовых нокдаун-экспериментах [Sanchez-Garcia и др., 2014], далеко не всегда драйверные мутации находятся в кодирующих областях. Это подчеркивает потенциальную опасность сбоев в контроле экспрессии генов, вызванных изменением регуляторных областей.

Некоторые типы сайтов связывания систематически разрушаются соматическими мутациями, что может отражать положительный отбор соответствующих геномных вариантов [Khurana и др., 2013; Melton и др., 2015]. В то же время для других факторов транскрипции сайты связывания избегают мутационных изменений [Khurana и др., 2013], однако достоверность отрицательного отбора ставилась под сомнение [Melton и др., 2015].

Для белок-кодирующих последовательностей давление отбора изучается на частотах синонимичных и несинонимичных замен [Ostrow и др., 2014]. Для некодирующих последовательностей оценка давления отбора может быть получена

на основании функциональной аннотации геномных вариантов. В частности, аннотация регуляторных районов может быть получена с помощью мотивов, для транскрипционных регуляторных районов – с помощью мотивов сайтов связывания факторов транскрипции: замены в консервативных (коровых) и вырожденных (фланкирующих) позициях в заметной степени аналогичны синонимичным и несинонимичным заменам в кодонах.

В описываемой работе [Vorontsov и др., 2016] мы использовали полногеномные данные о мутагенезе в различных типах раковых клеток [Alexandrov и др., 2013], чтобы идентифицировать соматические мутации, изменяющие сайты связывания конкретных факторов транскрипции, и оценить давление отбора. Для простоты изложения в ряде случаев под мотивом в этом разделе понимается набор сайтов связывания, соответствующих предсказаниям мотива в геноме.

5.2.5.1. Оценка давления отбора на мутации в сайтах связывания факторов транскрипции

Для изучения давления отбора на регуляторные области мы использовали данные по мутациям в различных типах раковых клеток, сгруппированных по ткани [Alexandrov и др., 2013]. Из полного набора мутаций мы выбрали только изменения в потенциальных регуляторных областях (в интронах и промоторах), которые составляли примерно половину общей выборки. Затем мы картировали потенциальные сайты связывания факторов транскрипции (используя модели HOCOMOCO) в небольших окнах, центрированных на мутациях, и рассматривали предсказания сайтов как перекрывающиеся с мутациями, так и находящиеся в окрестности (но не далее 10 п.н.). С одной стороны, это позволило отличить критические замены в ключевой коровой области сайтов от слабо-значимых или незначимых замен во фланкирующих позициях и ближайшей окрестности сайтов. С другой стороны, использование локальных окон позволило исключить влияние глобальной неравномерности мутагенеза в различных районах генома.

Возможные изменения аффинности сайтов оценивались для мутаций по сравнению с аллелями зародышевой линией с помощью PERFECTOS-APE [Vorontsov и др., 2015]. Учитывались оба направления изменения аффинности: уменьшение (ухудшение или разрушение сайта связвания) и увеличение (создание или улучшение сайта связывания, вызванное мутацией). Для оценки давления отбора

мы сравнивали наблюдаемые частоты мутаций, значительно меняющих предсказанную аффинность, с ожидаемыми частотами, оцененными по контрольным данным.

Для надежности было собрано два типа контрольных данных: (1) «перемешанный» контроль, состоящий из последовательностей со случайно переставленными нуклеотидами. Похожие решения использовались и ранее [Melton и др., 2015], но мы дополнительно сохраняли мутационный контекст (исходный нуклеотид, мутацию, и два нуклеотида, фланкирующих мутировавшую позицию). (2) геномный контроль, собранный из случайно сэмплированных сегментов промоторов и интронов, не перекрывающихся с окнами, центрированными на реальных мутациях.

Чтобы учесть систематический вклад мутационных подписей, характерных для различных типов раковых клеток, см. **Рисунок 43**, для каждого фактора транскрипции предсказанные сайты связывания в контрольных данных

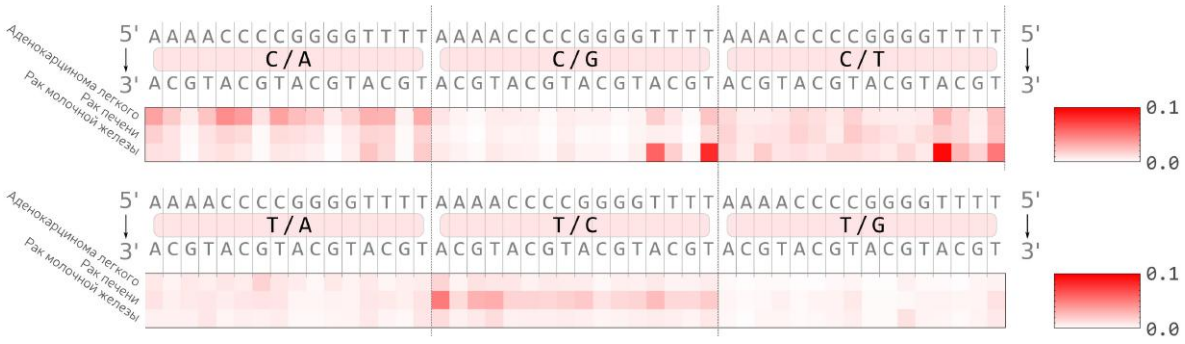


Рисунок 43. Относительные частоты мутаций в некодирующих районах геномов для трех типов рака с наибольшим числом детектированных соматических мутаций (аденокарцинома легкого, рак печени, рак молочной железы).

Мутации сгруппированы по типу замены, 5’ и 3’ нуклеотиды контекста показаны в лексикографическом порядке. Рисунок адаптирован из работы [Vorontsov и др., 2016].

сэмплировались, чтобы сделать распределение мутационных контекстов соответствующим наблюдаемому в каждом типе раковых тканей.

5.2.5.2. Давление отбора на мутации в регуляторных районах ограничено и требует больших выборок для обнаружения

Для каждого мотива мы оценили давление отбора на соматические мутации в предсказанных сайтах связывания. Величину давления мы условно определили как отношение наблюдаемой (для реальных мутаций) и ожидаемой (из контрольных данных) частот изменений аффинности. Результирующие значения попадают в интервалы около 0.9-0.95 (отрицательный отбор) и 1.05-1.1 (положительный отбор), и слабо отличаются при раздельном рассмотрении случаев уменьшения и увеличения аффинности.

Поскольку получаемые величины достаточно близки к 1 (а 1 соответствует совпадению наблюдаемой и ожидаемой частот мутаций в сайтах), для оценки статистической значимости потребовался большой объем данных. В частности, мы были вынуждены формировать большие контрольные выборки, превышающие объем выборок раковых мутаций в несколько раз (до десятков раз при исследовании конкретных типов рака с малым числом известных мутаций).

5.2.5.3. Мутации, изменяющие аффинность сайтов связывания, находятся под давлением отбора

Среди факторов транскрипции, сайты которых в результате мутаций теряют аффинность значительно чаще, чем ожидается, присутствовали члены семейств AP2 и C/EBP. Обогащение мутациями для этих сайтов находили и ранее [Khurana и др., 2013; Melton и др., 2015]. Также обогащены «повреждающими» мутациями оказались сайты связывания семейств SP и KLF. Интересно, что мутации в сайтах связывания других белков, в частности, членов семейства ETS, чаще, чем ожидается, вызывают усиление аффинности (т.е. создание сайта). Вхождения мотивов, затронутые мутациями чаще, чем ожидается, находятся под положительным отбором.

Для значительно более широкого спектра мотивов мутации, приводящие к изменению аффинности сайтов, избегаются в различных типах рака, см. **Рисунок 44**. В частности, мутаций избегают сайты связывания факторов транскрипции, принадлежащих классу ядерных рецепторов. Кроме того, под отрицательным

отбором найдены сайты белков семейств HOX и FOX, которые находятся под отбором и в нормальных клетках [Vernot и др., 2012]. Для типов рака с ограниченным числом известных соматических мутаций, статистическая значимость наблюдаемых эффектов позволяла выявить давление отбора только для отдельных мотивов. В то же время, эти «синглетоны» в ряде случаев принадлежали тем же семействам, которые были найдены под действием отрицательного отбора для типов рака с большими выборками мутаций.

5.2.5.4. Локализация соматических мутаций связана с информационным содержанием мотива

Частоты соматических мутаций существенно зависят от локального контекста последовательности. Собственный контекст мотива может интерферировать с мутационной подписью конкретного типа рака и неявно переопределять ожидаемые частоты мутаций в конкретных позициях. В то же время, можно ожидать, что давление отбора будет сильнее выражено для позиций с высоким информационным содержанием, поскольку замены нуклеотидов в них сильнее влияют на аффинность [Berg, 1987].

Для того чтобы сравнить частоты мутаций в различных позициях мотивов мы выровняли предсказания сайтов связывания в окнах, центрированных на мутациях с аллелями зародышевой линии. Затем позиционная плотность мутаций оценивалась путем нормализации частот мутаций в каждой позиции на полное число окон. На **Рисунке 45** показано распределение замен во вхождениях мотива AP2A (часто повреждаемых мутациями) и ESR1 (избегают мутагенеза) для рака молочной железы. Нормализованные частоты замен отображены параллельно лого-визуализации мотива: хорошо видно, что для мотива AP2A обогащена мутациями колонка G(+4). Известно, что в геноме рака молочной железы контекст 5’-TGA-3’ (5’-TCA-3’ на обратной комплементарной цепи) является высоко мутагенным. Это согласуется с тем, что мутации в позиции G(+4) превалируют по сравнению с остальными позициями, в том числе, в контрольных данных. Но при этом в геноме рака молочной железы частота мутаций G(+4) превышает контрольную в 1.5 раза.

Контрастная картина наблюдается в мажорной позиции C(+4) в контексте TCA мотива ESR1: мутации наблюдаются существенно реже по сравнению с любым из контролей. Не менее наглядно сравнение следующих двух боксов TGA: первый

центрирован на G(+10) и соответствует одинаковой частоте замен в раковых геномах по сравнению с контролями. Второй бокс, центрированный на G(+15), имеет более низкое информационное содержание и, вероятно, менее важен для связывания ESR1. В этой позиции аккумулируется значительно большее число соматических мутаций.

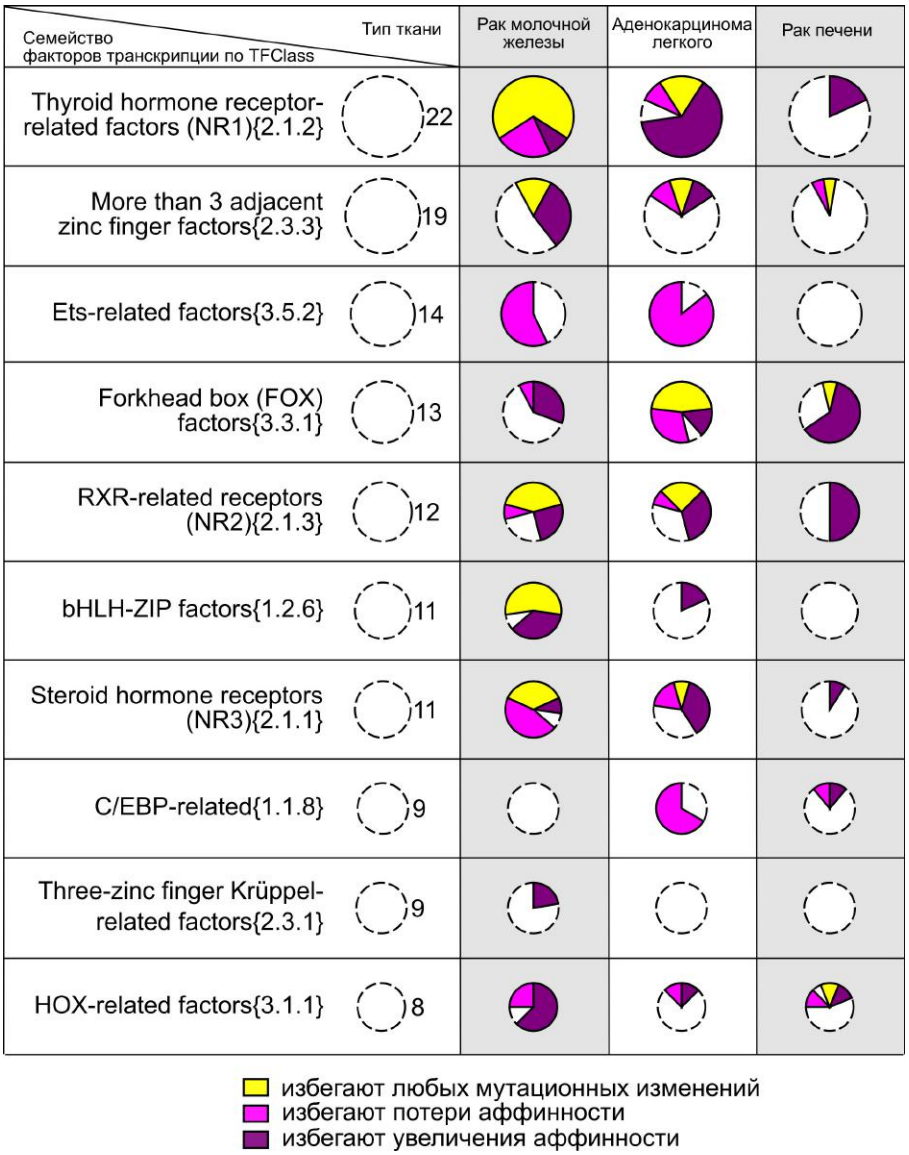


Рисунок 44. Мотивы факторов транскрипции, избегающие мутационных изменений в различных типах рака.

Размеры круговых диаграмм соответствуют числу мотивов семейства (по HOCOMOCO). Доли на круговых диаграммах соответствуют мотивам семейства, для которых наблюдается эффект: (*желтый*) мотивы избегают любых мутационных изменений, (*светло-фиолетовый*) мотивы избегают потери аффинности, (*темно-фиолетовый*) мотивы избегают роста аффинности. Семейства именованы в соответствии TFClass. Рисунок адаптирован из работы [Vorontsov и др., 2016].

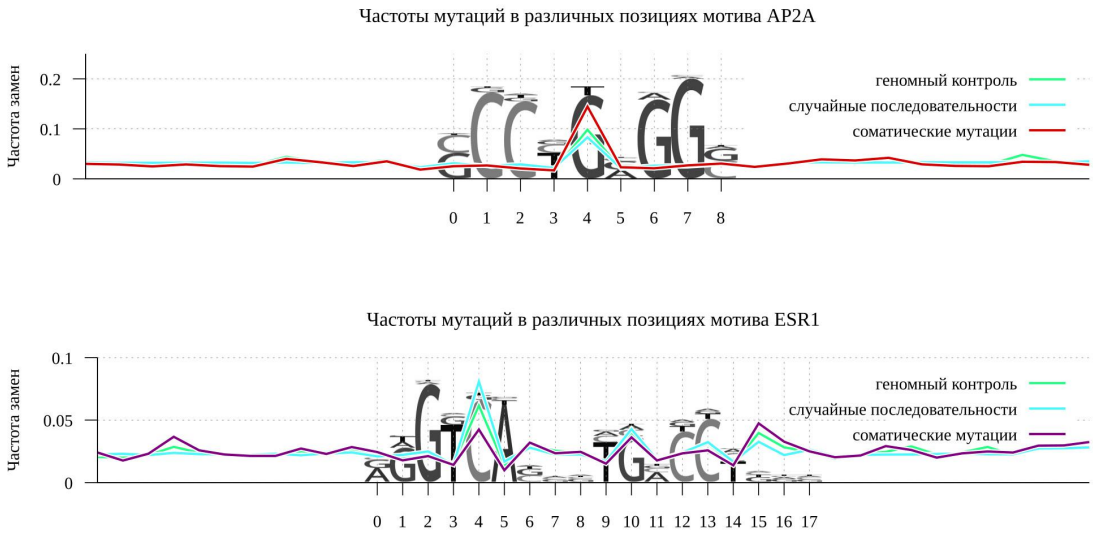


Рисунок 45. Относительное положение мутаций в геноме рака молочной железы по отношению к мотивам AP2A (*верхняя панель*) и ESR1 (*нижняя панель*).

Ось Y показывает относительную долю окон, центрированных на мутациях, содержащих вхождения мотивов для аллеля зародышевой линии. Ось X соответствует относительному положению мутаций. Цвета линий: (*красный*) соматические мутации мотива AP2A, (*фиолетовый*) соматические мутации мотива ESR1, (*голубой и зеленый*) контроли. Рисунок адаптирован из работы [Vorontsov и др., 2016].

5.2.5.5. Давление отбора на мутации в мотивах сильнее выражено в районах, доступных для эндонуклеазы

Прямое компьютерное предсказание сайтов связывания на основе анализа мотивов давно подвергается критике, достаточно сложно отличить настоящие сайты связывания от ложноположительных предсказаний без использования дополнительных экспериментальных данных. Чтобы увеличить достоверность предсказаний, мы рассматривали поднаборы мутаций, соответствующие ДНКазо-доступным участкам [Thurman и др., 2012] в промоторах и интронах, рассматривая мутационные профили рака молочной железы и аденокарциномы легкого. Данные по ДНКазо-доступности объединялись по близким клеточным линиям и родственным нормальным клеткам [Polak и др., 2015]. Абсолютное число мутаций и предсказаний сайтов связывания уменьшилось: фильтрованные наборы составляли порядка 30%/90% для рака молочной железы/аденокарциномы легкого, что уменьшило статистическую значимость. Тем не менее, основные наблюдения для ДНКазо-доступных районов совпали с полученными ранее для полных геномов, в частности,

отрицательный отбор был обнаружен для мотивов семейства FOX и нескольких семейств факторов, принадлежащих классу ядерных рецепторов, а члены семейств AP2 и C/EBP были найдены под положительным отбором, направленным на разрушение сайтов связывания. При детальном рассмотрении мотивов, для которых был обнаружен статистически значимый эффект отбора на полном наборе мутаций, выяснилось, что результаты по сравнению со «случайным» контролем в среднем сохраняются. «Геномный» же контроль показал значительно меньшую частоту мутационных изменений аффинности, что позволяет говорить о большем давлении отрицательного отбора на ДНКазо-доступные районы.

5.2.5.6. Обсуждение представленных результатов

Схожие мотивы демонстрируют схожие эффекты

Факторы транскрипции, принадлежащие одному структурному семейству, имеют похожие мотивы связывания, что осложняет или делает невозможным различение членов семейств на уровне последовательности. Именно поэтому мы фокусировались на наблюдениях, воспроизводимых для разных членов одного семейства, считая, что это позволяет избегать эффектов, обнаруженных для одного конкретного (возможно неточного) мотива. Дополнительный стабилизирующий фактор нашего анализа – использование мотивов, построенных по данным различных экспериментальных методов [Kulakovskiy и др., 2013a].

Использование двойного контроля выделяет наиболее достоверные эффекты

Использование геномного контроля в дополнение к «случайному перемешиванию» нуклеотидов необходимо: вхождения различных мотивов в регуляторных областях существенно неслучайны, например вхождения мотивов семейства SP коррелируют с составом CpG-островков, а сайты TBP – с мотивами связывания нуклеосом [22]. Регулярные особенности последовательностей, в том числе композитные элементы сайтов связывания, существенно влияют на локальный контекст и разрушаются при перемешивании букв. «Геномный» контроль давал в среднем более консервативные оценки на силу давления отбора, но итоговые наборы мотивов под отбором (детектируемые с помощью геномного и случайного контролей) существенно пересекались. Интересно, что наиболее согласованные результаты были получены для отрицательного отбора, а мотивы, потенциально попадающие под действие

положительного отбора, часто были значимы только по сравнению с одним контролем из двух. Например, сайты связывания мастер-регулятора ответа на гипоксию HIF-1 были найдены под сильным положительным отбором в обе стороны изменения аффинности по сравнению со случайным контролем, но все эффекты оказались незначимы при использовании геномного контроля. Мы не смогли предложить однозначной интерпретации этого наблюдения и исключили несогласованные между контролями результаты из рассмотрения.

Биологическая интерпретация

Сайты связывания, находящиеся под положительным отбором, соответствуют факторам транскрипции, которые ранее изучались в связи со злокачественной трансформацией клеток. Например, частое увеличение аффинности наблюдается для сайтов белков семейства C/EBP, для которых было показано участие в злокачественной трансформации эпителиальных клеток [Zahnow, 2009]. Сайты связывания GABP, члена ETS-семейства, создаются мутациями в промоторе теломеразы TERT и ассоциированы с развитием нескольких типов рака [Bell и др., 2015]. Мы обнаружили положительный отбор созданных сайтов ETS, в то время как потеря аффинности для них оказалась под отрицательным отбором. Белки семейства FOX, для которых сайты связывания защищены отбором от изменения аффинности в любую сторону, также ранее изучались в связи с онкогенезом [Myatt, Lam, 2007].

Важно понимать, что наше определение давления отбора имеет смысл для «ансамблей» сайтов связывания, которые изменяются мутациями чаще или реже, чем ожидается. Отдельный сложный вопрос – действие давления отбора на отдельные сайты связывания или мутации в локусах конкретных генов. Потенциально, мутагенез конкретных сайтов может выбиваться из общего тренда, выделенного для полного ансамбля мотивов белкового семейства. Поиск таких специфических случаев может помочь выявлению критических участков регуляторных сетей.

5.2.5.7. Методические замечания

Исходные данные по мутагенезу

Мы использовали опубликованный каталог мутаций, определенных с помощью полногеномного секвенирования 507 образцов для 10 типов рака (сгруппированных по тканям, [Alexandrov и др., 2013]). Наибольшее число мутаций было детектировано

для образцов рака молочной железы, рака печени и аденокарциномы легкого (1-2 замены на 5 тысяч п.н., мы рассматривали только однонуклеотидные варианты). Используя аннотацию генов Ensembl мы выбрали мутации в интронах и потенциальных промоторах и энхансерах рассматривая интервалы [-5000;+500] п.н. где 0 соответствует аннотированному сайту инициации транскрипции. Мутации в экзонах были исключены из рассмотрения. В качестве основного объекта использовались геномные «окна» [-50; +50] п.н., центрированные на мутациях. Перекрывающиеся окна рассматривались независимо, исключались дублирующиеся окна (мутации в повторяющихся сегментах генома) в рамках одного образца, рецидивные мутации в конкретной позиции в разных образцах считались независимыми наблюдениями.

Оценка изменения аффинности сайтов связывания

Мы считали значимым изменение *P*-значения как минимум в 4 раза: в идеальном случае это аналогично замене полностью консервативной колонки мотива (с единственным допустимым нуклеотидом) на случайную колонку с частотами нуклеотидов, соответствующими ожидаемым из фонового распределения (например, равномерными). Для предсказаний сайтов связывания использовали 278 мотивов A/B/C-качества из коллекции HOCOMOCO v9. Для предсказания сайтов и оценки *P*-значений использовали PERFECTOS-APE (см. «Материалы и методы»), с допустимой локализацией сайта в ближайшей окрестности мутирующих позиций (вплоть до 10 п.н.). Единый порог на *P*-значения мотивов составлял 0.0005, как минимум для одного из двух тестируемых вариантов последовательности. Чтобы получить качественную оценку на реальную частоту ложноположительных предсказаний мы использовали данные EMSA для FoxA2. В них мотив из HOCOMOCO v9 при фиксированном пороге, соответствующем *P*-значению в 0.0005, показывал долю ложноположительных предсказаний около 17%. К сожалению, эта оценка достаточно грубая, поскольку сделана на 64 олигонуклеотидах (41 прошедших верификацию как сайты связывания и 23 «не-сайтах»), которые исходно выбирались как наиболее спорные с точки зрения компьютерного распознавания (см. соотв. секцию в разделе «Результаты и обсуждение»).

Оценка статистической значимости

Чтобы для конкретного типа рака оценить значимость отличий наблюдаемой и ожидаемой частот изменения аффинности сайтов, определяемых конкретным мотивом, мы использовали точный текст Фишера на четырехпольных таблицах сопряженности (значительное увеличение или уменьшение аффинности по сравнению с незначительным; реальные мутации по сравнению с контролем). Тест повторялся для каждого мотива дважды, с использованием геномного и случайного контролей. В дальнейшем учитывались только случаи, прошедшие порог в 0.05 по статистической значимости теста Фишера для обоих контролей после FDR-поправки [Benjamini, Hochberg, 1995] на множественное тестирование 278 мотивов.

Оценка силы отбора

Частота событий уменьшения аффинности связывания (повреждения сайта вследствие мутаций) подсчитывалась как отношение числа центрированных на мутациях геномных окон со значительным снижением аффинности к общему числу окон с предсказаниями сайтов для вариантов зародышевой линии (раздельно для каждого мотива). Аналогично, частота событий увеличения аффинности связывания подсчитывалась как число окон со значительным увеличением аффинности (создание сайта *de novo*) подсчитывалась по отношению к окнам с предсказаниями для мутировавшего аллеля.

Абсолютные частоты существенно зависят от мутационной подписи конкретного типа рака и конкретного мотива. Поэтому, в качестве нормализованной оценки мы подсчитывали отношение наблюдаемой частоты для реальных мутаций и ожидаемой частоты, определенной в контроле, сэмплированном для подбора аналогичной представленности мутационных контекстов. Отношение частот мы называем оценкой давления отбора, значения больше 1 соответствуют положительному отбору, значения меньше единицы – отрицательному.

5.2.5.8. Заключение по разделу

Нам удалось продемонстрировать, что сайты связывания ряда факторов транскрипции в геномах раковых клеток мутируют значительно реже, чем ожидается. Избегание мутаций свидетельствует об отрицательном отборе, поддерживающем определенные участки регуляторных сетей. Тот факт, что

избегаются и мутации, повреждающие сайты связывания, и мутации, потенциально способные создать новые сайты, позволяет утверждать, что наблюдаемая разница не вызвана «механистической защитой от мутагенеза», обеспечиваемой самими факторами транскрипции, занимающими соответствующие сайты на ДНК. Другое потенциальное объяснение, что статистические отклонения объясняются специфическими неучтенными особенностями мутационных подписей, также может быть исключено, поскольку многие сохраняемые семейства мотивов совпадают между различными типами рака. Более того, даже простой тест, сравнивающий частоты мутаций вокруг вхождения мотива и в самом вхождении (без учета изменения аффинности) выделяет те же семейства: мотивы ETS и AP2 в среднем чаще перекрываются с мутациями, а FOX и HOX избегают мутаций.

Дальнейший анализ сайтов связывания и потенциальных генов-мишеней должен помочь в выделении конкретных регуляторных путей, критических для онкогенеза.

5.2.6. Идентификация мотивов в промоторах проекта FANTOM5

Контролируемая экспрессия генов определяет развитие и пространственную организацию сотен типов клеток, составляющих организмы высших эукариот. Базовым этапом реализации генетической информации через экспрессию является транскрипция, специфичность которой, во-многом, и определяет клеточную идентичность. Проект FANTOM (Functional ANnoTation Of the Mammalian genome) был начат в институте RIKEN (Институт физико-химических исследований, Япония) в 2000 году под руководством Йошихида Хаяшизаки (Yoshihide Hayashizaki). Первичной целью проекта была функциональная аннотация неизученных на тот момент полноразмерных кДНК мыши, но последовательное развитие международного партнерства позволило на новых итерациях решать более амбициозные и масштабные задачи. Проект FANTOM5 [Forrest и др., 2014], в котором участвовала российская группа под руководством В.Ю. Макеева (ИОГен РАН), ставил своей целью получение полномасштабного атласа транскрипционной активности различных типов клеток, как иммортализованных клеточных линий, так и первичных клеток и нормальных тканей. Для решения этой задачи использовалась технология кэп-анализа экспрессии генов (CAGE) и метод высокопроизводительного секвенирования отдельных молекул Helicos (HeliScopeCAGE), который исключает

шаг ПЦР, и, как следствие, позволяет получить точные количественные оценки активности генов. В ходе проекта FANTOM5[50] было выявлено более двух сотен тысяч активных участков инициации транскрипции в геномах человека и мыши, большинство идентифицированных стартов транскрипции были специфичными для конкретных типов клеток или тканей. Роль российской группы заключалась в самостоятельной идентификации мотивов в промоторах, активных в различных типах клеток, и интеграции результатов идентификации мотивов, полученных другими членами консорциума. Такой анализ позволил провести независимую «ортогональную» верификацию ткань-специфичной активности промоторов путем сравнения найденных мотивов с известными, принадлежащими конкретным ткань-специфичным регуляторам и кроме того найти принципиально новые регуляторные паттерны.

5.2.6.1. *De novo* идентификация мотивов связывания

Систематическая идентификация мотивов в ткань-специфичных промоторах была проведена с помощью четырех инструментов: нашего метода ChIPMunk (процедура подробнее обсуждается ниже, было получено 1619/630 мотивов для человека и мыши, соответственно) и трех альтернативных методов, использованных членами консорциума, – DMF [Marchand, Bajic, Kaushik, 2011] (848/837 мотивов), ScanAll [Zantoni и др., 2007] (1370/1277) и HOMER [Heinz и др., 2010] (1426/692). Во всех случаях использовались специфичные для конкретного типа клеток или ткани поднаборы промоторов, непосредственно окружающих регионы инициации транскрипции.

Идентификация мотивов с помощью ChIPMunk

Для идентификации мотивов мы использовали проксимальные промоторы регионов инициации транскрипции, заданные как интервалы [-400 в 5’ область; +100 в 3’ область] вокруг CAGE-кластеров (регионов, определенных в ходе FANTOM5 по выраженной кластеризации 5’ концов картированных чтений, соответствующих 5’ концам транскриптов, отобранных за кэп).

ChIPMunk позволяет использовать априорные позиционные данные о положении мотива. Для каждой последовательности мы сгенерировали

[50] FANTOM5 – Functional Annotation of the Mammalian Genome. http://fantom.gsc.riken.jp/5/

трапециевидный профиль, равный нулю на внешних границах областей, и максимальный на протяжении CAGE-кластера. Высота трапеции (вес профиля) всякий раз принималась равной нормализованной активности региона инициации транскрипции в конкретном типе клеток. Эта процедура позволяла сфокусировать поиск мотивов на активно экспрессируемых промоторах, и отдать предпочтение мотивам, вхождения которых были локализованы в непосредственной близости от участков инициации. Мы использовали два вида фонового распределения – равномерное и усредненное по промоторам, активным в конкретной ткани. Расширение ChIPHorde использовалось для итеративного поиска (до 3х мотивов с поли**N**-маскированием результатов каждого предыдущего шага). Максимальная длина мотива была ограничена 12 п.н. (как медианная длина сайтов связывания), для улучшения сходимости использовался однобоксовый вариант «формы» мотива.

Результаты идентификации мотивов были отфильтрованы: мотивы были отсортированы по шагу маскирования (отдавалось предпочтение ранним шагам), затем по длине (считая более длинные более информативными), затем по весу выравнивания (больший вес соответствовал большей суммарной экспрессии соответствующих промоторов). Наконец использовался жадный алгоритм фильтрации ранжированного списка мотивов, а именно для наилучшего (первого) мотива по списку удалялись все похожие на него, и эта процедура повторялась для всех членов списка. Сходство мотивов оценивали с помощью MACRO-APE на уровне *P*-значений 0.0005 и пороговым значением 0.2 для сходства по Жаккару.

5.2.6.2. Оценка новизны мотивов

Для идентификации мотивов *de novo* использовались характерно разные подходы: учет позиционной информации (ChIPMunk), дифференциальный поиск мотивов по сравнению с контрольной выборкой и словарными затравками (DMF, HOMER), идентификация композитных элементов с фиксированным спейсером (ScanAll). Ключевые вопросы: насколько хорошо найденные мотивы описывают множество известных регуляторных сигналов и удалось ли идентифицировать принципиально новые сигналы, характерно отличающиеся от всех известных. Для ответа на эти вопросы использовались актуальные на тот момент релизы коллекций известных мотивов: HOCOMOCO (v9, 426 мотивов); HOMER (138 мотивов, основанных на ChIP-Seq данных), JASPAR (130 мотивов набора CORE VERTEBRATE);

SwissRegulon (190 мотивов); UniPROBE (413 мотивов, полученных по данным белок-связывающих микрочипов); и, дополнительно, «регуляторный лексикон» ENCODE LEXICON (683 мотива, из них 289 были определены как потенциально новые) [Neph и др., 2012], полученный в ходе проекта ENCODE на основе анализа полногеномного футпринтинга ДНКазой I.

Чтобы оценить сходство 8699 *de novo*-мотивов с известными, для каждой коллекции известных мотивов был запущен TomTom [Gupta и др., 2007]. *De novo*-мотив считался похожим на известный, если *E*-значение TomTom не превышало 0.5 (т.е. ожидалось менее одного похожего мотива при подаче случайного мотива на поиск в известной коллекции). Для 7478 из 8699 *de novo*-мотивов были найдены известные аналоги, оставшийся 1221 мотив потенциально мог соответствовать новым регуляторным сигналам.

Чтобы оценить, насколько хорошо покрыто множество известных мотивов, TomTom раздельно запускали «в обратную сторону» для поиска каждого из известных мотивов, используя объединенный набор *de novo*-мотивов как единую коллекцию. Прямое объединение всех *de novo* мотивов опасно, поскольку TomTom при поиске масштабирует *E*-значение в соответствии с размером коллекции и это не позволяет корректно выбрать порог для экстремально большой коллекции. Поэтому, использовался двухэтапный подход, когда сначала для каждого известного мотива был выбран наиболее похожий *de novo*-вариант, таким образом, отобрали всего 1105 *de novo*-мотивов. Это позволило безопасно повторить процедуру для *de novo* мотивов «в обратную сторону» с использованием того же фиксированного порога 0.5 на *E*-значение. В итоге для 90% известных мотивов были успешно обнаружены аналоги, выявленные *de novo*. То есть, лишь около 10% известных мотивов не удалось выявить при систематическом анализе промоторов и обратно, лишь 10% выявленных «потенциально новых» мотивов значительно отличались от известных. По сути это означает, что общий репертуар регуляторных мотивов млекопитающих сегодня уже практически изучен.

5.2.6.3. Выявление принципиально новых мотивов

TomTom спроектирован для оценки вероятности случайности нахождения мотива в заданной коллекции, но для конкретного мотива и различных коллекций это

осложняет интерпретацию *E*-значений, которые зависят от размера и состава (степени избыточности) коллекции. Процедура, использующая TomTom, позволяет хорошо оценить общую степень похожести наборов мотивов между собой. Однако, остается открытым вопрос, насколько уникальными и разнообразными являются мотивы, для которых не нашлось похожих примеров в прямом (*de novo* против известных) или обратном (известные против *de novo*) сравнениях.

Для решения этой задачи мы провели иерархическую кластеризацию (UPGMA) для объединенного набора всех известных и *de novo*-мотивов, основываясь на сходстве Жаккара, которое подсчитывалось с помощью MACRO-APE для *P*-значения 0.0005. Ветви иерархического дерева итеративно обрезались на различной высоте (используя порог на длину связи), и мотивы на одной ветви объявлялись кластером. Мы оценили зависимость числа аннотированных кластеров от порога кластеризации: аннотированным кластером называли кластер *de novo*-мотивов, которому принадлежал хотя бы 1 известный мотив. Характерный максимум числа аннотированных кластеров достигался на порогах сходства Жаккара около 0.05 (218 аннотированных кластеров, на длине связи в дереве равной 0.9586). Затем мы независимо кластеризовали 1221 «потенциально новых» мотивов, ранее отфильрованных TomTom, и использовали фиксированный порог на длину связи в дереве для получения итогового набора из 172 кластеров. Для всех мотивов кластера подсчитывалась корреляция между предсказаниями сайтов связывания и активностью промоторов в различных клетках, мотив с наивысшей корреляцией выбирался как репрезентативный. Для 169 кластеров удалось выбрать репрезентативный мотив, вхождения которого были скоррелированы с экспрессией промоторов, и для 37 мотивов наблюдаемая корреляция оказалась значимой с $P < 0.05$ (1000 случайных перемешиваний колонок весовых матриц и поправка Бонферрони на множественное тестирование 169 мотивов). Релевантную функциональную роль этих мотивов подтверждает и анализ обогащенности целевых промоторов, содержащих вхождения, терминами генной онтологии [McLean и др., 2010], и неравномерность позиционного распределения вхождений относительно региона инициации транскрипции. К сожалению, пока трудно сказать, соответствуют ли новые мотивы посадке неизученных факторов транскрипции, или другим регуляторным механизмам. Схема процедуры представлена на **Рисунке 46**.

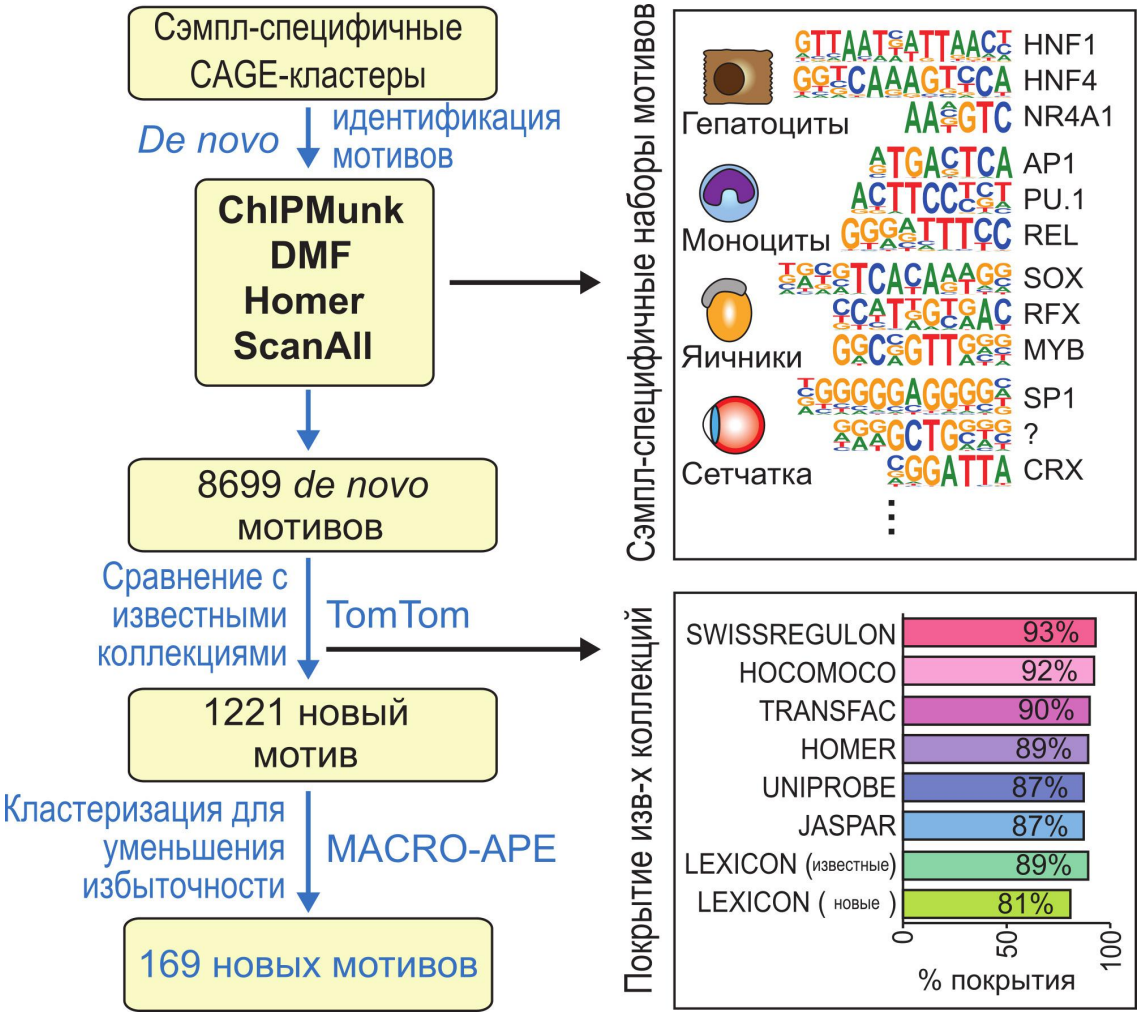


Рисунок 46. Схема идентификации и аннотации мотивов на основе промоторов, активных в различных биологических образцах (тканях, первичных клетках, иммортализованных клеточных линиях) и выявленных с помощью технологии CAGE. Рисунок адаптирован из материалов консорциума FANTOM5 [Forrest и др., 2014]. Интересно, что лексикон мотивов, определенных ENCODE в ДНКазо-доступных областях и не проаннотированных как известные, наихудшим образом покрывается мотивами из CAGE-промоторов.

5.2.7. Колокализация сайтов связывания факторов транскрипции и CpG-светофоров

Описываемый в этом разделе проект был проведен Ю. Медведевой в KAUST (King Abdullah University of Science and Technology, Саудовская Аравия) в рамках партнерства FANTOM5 в составе международной группы под руководством В. Байича. В ходе работы была использована наша коллекция моделей сайтов связывания (HOCOMOCO), непосредственно как для полногеномного предсказания сайтов, так и при построении более сложных моделей, учитывающих удаленные зависимости между позициями (Remote-Dependency Model, RDM, разработаны в группе В. Байича). Анализ коровых и фланкирующих позиций в сайтах мы проводили на основании исходных ПВМ из HOCOMOCO.

5.2.7.1. Метилирование ДНК и активность промоторов млекопитающих

Метилирование ДНК является одной из наиболее изученных эпигенетических модификаций, которая выполняет регуляторную функцию в разнообразных нормальных и патологических процессах: онтогенетическом развитии [Borgel и др., 2010; Jaenisch, Bird, 2003], поддержании клеточной идентичности [Oda и др., 2006], дифференцировке [Farthing и др., 2008], поддержании плюрипотентности [Tomazou, Meissner, 2010], старении [Christensen и др., 2009], развитии онкологических [Dawson, Kouzarides, 2012] и нейродегенеративных заболеваний [Kwok, 2010].

В дифференцированных клетках млекопитающих метилирование цитозина в большинстве случаев происходит в составе CG-пар, в свою очередь, до 90% цитозинов в CpG-контексте являются метилированными [Ehrlich и др., 1982]. Метилирование CpG является стабильным по отношению к типу клеток, что позволяет говорить о его систематическом вкладе в генную регуляцию.

Для изучения метилирования в геномном масштабе часто используется секвенирование с бисульфитной конверсией ДНК [Laird, 2010]. Информация о метилировании отдельных CpG-сайтов традиционно используется только для оценки общего метилирования окружающего района, что согласуется с кластеризацией неметилированных CpG (в CpG-островках) и метилированных CpG (в повторах). Существует гипотеза о различной регуляторной роли метилирования в CG-богатых и CG-бедных участках генома [Schubeler, 2012]. В то же время, отдельные CpG могут

быть метилированы и в CpG-островках [Lister и др., 2009], и, более того, конкретный дифференциально метилированный CpG-сайт может влиять на транскрипцию гена (пример ESR1 обсуждается в работе [Fürst и др., 2012]). Вопрос о систематической регуляторной роли отдельных CpG-сайтов на уровне транскрипции долгое оставался открытым.

Метилирование промоторов тесно связано с ингибированием транскрипции [Walsh, Bestor, 1999], но конкретные молекулярные механизмы остаются неясными. Во многих работах показана отрицательная корреляция метилирования промоторов и экспрессии соответствующих генов [Jones, Takai, 2001; Shen и др., 2007]. В то же время, транскрипция некоторых генов не зависит от метилирования [Bird, 1984], в частности, промоторы с низким GC-составом обычно метилированы, но вопреки этому могут сохранять активность [Eckhardt и др., 2006; Weber и др., 2007], возможно, за счет компенсаторного действия энхансеров [Boyes, Bird, 1992]. Иногда метилирование нужно для активации транскрипции и положительно скоррелировано с экспрессией генов [Rishi и др., 2010]. Таким образом, направление (активация или игнибирование) и механизмы действия метилирования промоторов на экспрессию генов также требуют изучения.

Одним из объяснений ассоциации метилирования конкретных CpG-сайтов с активностью промоторов могла бы быть их прямая механистическая роль, например, локализация в сайтах связывания факторов транскрипции и изменение аффинности связывания при метилировании конкретных CpG-позиций.

Существует масса несистематических данных о влиянии метилирования ДНК на связывание факторов транскрипции [Choy и др., 2010; Perini и др., 2005; Tate, Bird, 1993]. В то же время, большинство экспериментально определенных сайтов связывания находятся в неметилированном состоянии [Mukhopadhyay и др., 2004; Thurman и др., 2012], и статус метилирования напрямую определяет возможность связывания регулятора [Gebhard и др., 2010; Straussman и др., 2009]. Для сайтов CTCF различия в метилировании были обнаружены даже для различных позиций сайтов связывания [Wang и др., 2012a]. Наконец, метилирование цитозина может в некоторых случаях привлекать регуляторы транскрипции, как активаторы [Chatterjee, Vinson, 2012], так и репрессоры; например белки семейства MBD (Methyl-binding

domain), связывая метилированые CpG-сайты, запускают конденсацию хроматина и репрессируют активность генов-мишеней [Nan и др., 1998].

Возможна и альтернативная модель, в которой модификации хроматина уменьшают доступность сайтов связывания для факторов транскрипции [Feldman и др., 2006; Strunnikova и др., 2005]; в этом случае метилирование ДНК является не причиной снижения активности гена, а маркером, аккумулируемым при отсутствии связывания факторов транскрипции [Thurman и др., 2012; Wang и др., 2012a]; или результатом прямого привлечения метилтрансферазы белками-репрессорами семейства Polycomb [Viré и др., 2006]. Эта модель поддерживается данными об обратной зависимости среднего метилирования сайтов связывания от экспрессии соответствующих факторов транскрипции [Thurman и др., 2012].

Чтобы прояснить, как и какие именно факторы транскрипции соответствуют озвученным моделям, был проведен совместный систематический анализ позиционных данных по метилированию, ткань-специфичной активности промоторов и предсказанных сайтов связывания факторов транскрипции.

5.2.7.2. Определение CpG-светофоров

В данном проекте были использованы данные ENCODE [Dunham и др., 2012] по метилированию ДНК, полученные секвенированием с ограниченной бисульфитной конверсией (reduced representation bisulfite sequencing, RRBS) [Gu и др., 2011; Meissner и др., 2005]. Это позволяет выяснить локализацию метилированных и неметилированных цитозинов в CG-богатых районах генома, обогащенных тетрануклеотидами CCGG, в первую очередь в промоторах и CpG-островках. Для оценки активности промоторов в различных типах клеток использовали данные проекта FANTOM5 [Forrest и др., 2014]. Было выбрано 50 образцов ENCODE (36 нормальных и 14 раковых), для которых были вручную сопоставлены 137 образцов FANTOM5 (с усреднением близких образцов или реплик). Для оценки связи метилирования и экспрессии подсчитывался коэффициент корреляции Спирмана, значимость значения которого оценивалась с помощью преобразования Фишера. Среди более чем 200 тысяч проанализированных CpG-сайтов, менее процента показывали положительную корреляцию экспрессии с метилированием и более 15% (10% для нормальных тканей) – негативную корреляцию ($P < 0.01$). Такие CpG-сайты

были названы CpG-светофорами, которые схематично проиллюстрированны на **Рисунке 47**. Почти 80% светофоров локализовались в промоторах генов (-1500 п.н. в 5’ область и до +500 п.н. в 3’ область относительно CAGE-кластера).

Следующим шагом был анализ колокализации CpG-светофоров и сайтов связывания факторов транскрипции.

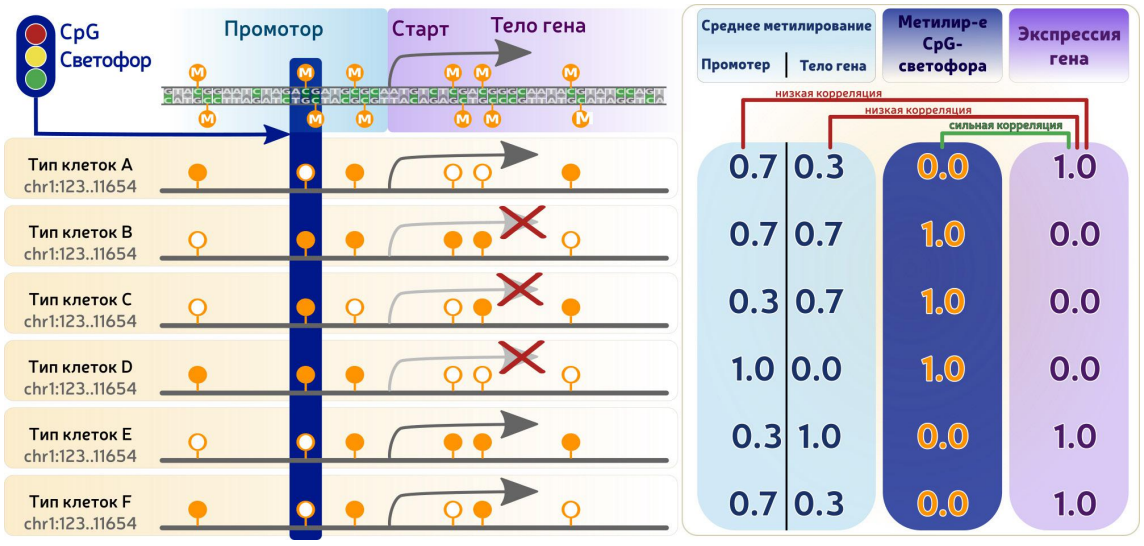


Рисунок 47. Схема CpG-светофора. Рисунок адаптирован по макету, любезно предоставленному Анной Лиозновой и Юлией Медведевой.

5.2.7.3. Сайты связывания факторов транскрипции избегают CpG-светофоров

Предположим, что CpG-светофоры не являются побочным эффектом от среднего метилирования неактивных промоторов. Тогда отрицательная ассоциация светофоров с экспрессией может быть вызвана «выключением» связывания конкретных факторов транскрипции через изменение локальной структуры ДНК.

Были выполнены полногеномные предсказания сайтов связывания с помощью моделей RDM (Remote-Dependency Model, модели с учетом удаленных зависимостей) и традиционных весовых матриц (используя стандартные пороги оценок по *P*-значениям 0.0005, оцененным с помощью MACRO-APE). Для RDM-предсказаний, перекрытие сайтов связывания с CpG-светофорами было обнаружено для 271 фактора транскрипции, и для 100 из них была обнаружена значимая недопредставленность CpG-светофоров в сайтах связывания (P-значение < 0.05, тест χ^2 с поправкой Бонферрони на множественное тестирование факторов транскрипции), и лишь один мотив (OTX2, предположительно артефактный) предпочитал значимо колокализоваться с CpG-светофорами. Для 36 типов нормальных клеток, рассмотренных отдельно, были получены похожие результаты: сайты связывания 35 факторов транскрипции избегали перекрытия с CpG-светофорами, значимое обогащение пронаблюдать не удалось. Распределение отношения «наблюдаемое-к-ожидаемому» для числа сайтов связывания, перекрывающихся с CpG-светофорами (оцененное по числу светофоров среди всех CpG-сайтов) показано на **Рисунке 48**. При рассмотрении предсказаний традиционных весовых матриц, сайты 270 из 279 факторов транскрипции избегали перекрытий с CpG-светофорами.

Метилирование цитозина может мешать связыванию конкретных факторов транскрипции (например, Sp1 [Macleod и др., 1994] или CTCF [Wang и др., 2012a]). Мы предполагали, что точечное метилирование может работать как глобальный регуляторный механизм, включающий или выключающий ключевые сайты связывания. Однако, результаты систематического анализа говорят об обратном: сайты связывания систематически избегают CpG-сайтов, метилирование которых скоррелировано с экспрессией. Можно предполагать, что это вызвано давлением отбора, исключающим систематическое выключение функциональных сайтов связывания с помощью метилирования.

Результаты были дополнительно проверены на прямых данных по связыванию CTCF, удалось подобрать ChIP-Seq для 22 из 50 использованных типов клеток. И вновь, наблюдения показали систематическое избегание CpG-светофоров.

Используя аннотацию UniProt мы раздельно проанализировали факторы транскрипции по фунциональным классам: активаторы и репрессоры. Выяснилось, что репрессоры сильнее избегают CpG-светофоров, чем активаторы; но общая картина осталась неизменной.

Наконец, мы проанализировали коровые (удовлетворяющие порогу «3 из 4» на дискретное информационное содержание, см. описание программы ChIPMunk в разделе «Материалы и методы») и фланкирующие позиции сайтов связывания. Выяснилось, что вероятность обнаружить CpG-светофор в коровой позиции сайта связывания еще ниже, чем в среднем, хотя статистическая оценка разницы, наблюдаемой между корами и фланкирующими позициями сайтов, незначима в силу малых выборок.

Таким образом, с одной стороны, в работе удалось выявить существование CpG-светофоров, единичных CpG-сайтов, значимо скоррелированных с экспрессией промоторов. С другой стороны, механизм этой связи остается неясным и, судя по полученным данным, не осуществляется через точечную модификацию аффинности сайтов связывания факторов транскрипции.

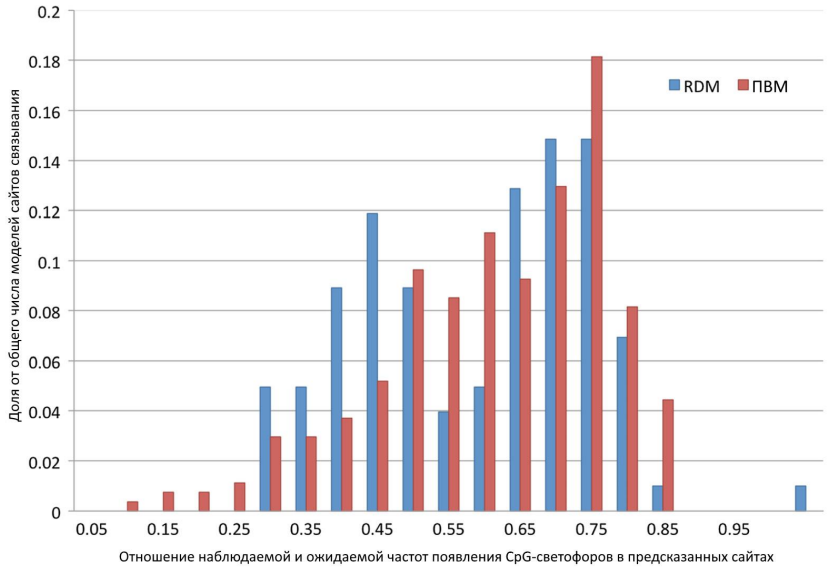


Рисунок 48. Распределение отношений ожидаемых и наблюдаемых частот попадания CpG-светофоров в сайты связывания различных факторов транскрипции. Рисунок адаптирован из работы [Medvedeva и др., 2014].

6. Заключение

В ходе работы нам удалось создать комплекс биоинформатических методов для анализа мотивов и взглянуть на локализацию и функцию мотивов в некодирующих районах геномов с точки зрения различных задач по изучению регуляции экспрессии генов высших эукариот. Можно смело прогнозировать, что развитие биоинформатических методов для анализа мотивов не закончится на текущем витке. Для ДНК-паттернов все еще не найден идеальный способ представления, пригодный для описания всего спектра экспериментальных данных и удобный на практике. В результате, традиционная упрощенная модель мотива – позиционно-весовая матрица – все еще остается наиболее популярной. В то же время, фокус исследователей смещается от локального анализа мотивов к использованию машинного обучения для достоверного полногеномного предсказания сайтов, активных в конкретном типе клеток: именно такая задача ставилась перед участниками соревнования DREAM-ENCODE 2016 по машинному обучению в биологии[51].

В этом есть прямой практический смысл: секвенирование продолжает дешеветь, но прямой экспериментальный анализ ДНК-белковых комплексов требует серьезных затрат помимо секвенирования. Всеобъемлющий эксперимент для всех сотен типов клеток человеческого организма вряд ли будет возможен в обозримом будущем, а необходимость в детальной регуляторной карте генома есть уже сегодня. Так, прочтение индивидуального генома для задач персонализированной медицины является решаемой задачей, в отличие от полноценной интерпретации индивидуальных геномных вариантов, в том числе, находящихся в регуляторных областях. К примеру, именно в промоторах и энхансерах находится более 60% казуальных вариантов, определяющих предрасположенность к аутоимунным заболеваниям [Farh и др., 2015].

Можно надеяться, что удивительный феномен индивидуальной невосприимчивости к серьезнейшим менделевским заболеваниям также будет объяснен вариантами регуляторных последовательностей [Chen и др., 2016]. Вряд ли анализ мотивов будет достаточен для понимания всех индивидуальных особенностей регуляции экспрессии, даже если ограничиться только сайтами связывания факторов

[51] биомолекула.ру: Мечту вызывали? http://biomolecula.ru/content/2064

транскрипции. Продвинуться в этом направлении несомненно поможет привлечение разносторонних данных, для уровня транскрипции важнейшую роль будет играть информация о доступности и организации хроматина [Deplancke и др., 2016].

Эйфория от быстрого прочтения геномных последовательностей сменилась волнующим осознанием бесконечной длины пути от фактического прочтения генома до функционального описания механизмов экспрессии генов. Сегодня мы обладаем спектром мощнейших, пусть и не идеальных, экспериментальных методов по анализу генной регуляции, порождающих данные невиданного ранее объема и сложности. В этом свете можно смело утверждать, что биоинформатика сыграет решающую роль в развитии молекулярной биологии и понимания механизмов регуляции экспрессии генов.

7. Выводы

1. Разработаны новые биоинформатические методы для анализа мотивов в нуклеотидных последовательностях: методы идентификации мотивов в форме моно- и динуклеотидных весовых матриц с использованием современных масштабных экспериментальных данных, основанных на высокопроизводительном секвенировании; метод сравнения мотивов на основании естественной меры сходства по Жаккару. Усовершенствован метод оценки различий регуляторного потенциала однонуклеотидных вариантов на основе *P*-значений мотивов. Методы реализованы в виде компьютерных программ с открытым исходным кодом и предоставлены в открытый доступ в сети Интернет.

2. Проведена идентификация и тестирование мотивов ДНК-белкового узнавания на основе опубликованных в открытом доступе результатов нескольких сотен масштабных экспериментов по изучению связывания ДНК факторами транскрипции как *in vivo* (ChIP-Seq), так и *in vitro* (HT-SELEX). Создана новая коллекция мотивов сайтов связывания факторов транскрипции мыши и человека, наиболее полная по сравнению с ранее опубликованными. Представленные в коллекции мотивы в форме позиционно-весовых матриц покрывают основные структурные семейства факторов транскрипции и наилучшим образом среди существующих источников представляют характерные паттерны, соответствующие участкам ДНК, связывающим факторы транскрипции. В представленную коллекцию впервые систематически включены расширенные динуклеотидные модели, учитывающие корреляции между соседними нуклеотидами. Коллекция предоставлена в открытый доступ в сети Интернет в виде интерактивной базы данных.

3. С помощью новых биоинформатических методов получен ряд важных результатов в регуляторной геномике эукариот:

(а) На основании систематического анализа ChIP-Seq данных выдвинута гипотеза о тройственном композитном OSN-элементе сайтов связывания OCT4-SOX2/NANOG, через который реализуется антагонизм OCT4 и NANOG в регуляции генов плюрипотентности.

(б) Выявлено, что действие фактора транскрипции Spi1 на экспрессию генов-мишеней в мышиной модели эритролейкемии зависит от локализации и контекста тандемных пар участков связывания: тандемные сайты чаще всего активируют экспрессию генов при локализации в промоторах, не пересекающихся с CpG-островками, и ингибируют экспрессию, если локализованы в энхансерах.

(в) Подтверждена предпочтительная локализация пиримидин-богатых мотивов в непосредственной окрестности 5’-концов мРНК-мишеней сигнального каскада mTOR с помощью совместного анализа данных рибосомного профайлинга и детальной карты регионов инициации транскрипции. Обнаружение протяженных регионов инициации транскрипции говорит в пользу альтернативной регуляции mTOR-мишеней на уровне трансляции в зависимости от транскрипционной активности промоторов.

(г) Установлено действие отрицательного отбора на мутации, возникающие в окрестности сайтов связывания факторов транскрипции в геномах раковых клеток. Показано, что мотивы ряда семейств факторов транскрипции избегают мутационных изменений, как повреждающих существующие сайты, так и направленных на создание новых, что свидетельствует о сохранении критических участков регуляторных сетей.

(д) В ходе проекта FANTOM5 успешно проведена идентификация мотивов в полногеномном каталоге промоторов, определенных в различных типах клеток, выявлены мотивы известных ткань-специфичных регуляторов и установлено, что разнообразие промоторных регуляторных сигналов практически полностью описывается каталогом известных регуляторных мотивов.

(е) Установлено систематическое избегание колокализации сайтов связывания факторов транскрипции и CpG-светофоров, конкретных CG-пар генома, дифференциальное метилирование которых скоррелировано с активностью близлежащих промоторов.

8. Публикации и доклады по теме диссертации

8.1. Статьи в рецензируемых международных журналах

вх. в список Web of Science и Scopus

1. (2017) A.M. Schwartz, D.E. Demin, I.E. Vorontsov, A.S. Kasyanov, L.V. Putlyaeva, K.A. Tatosyan, I.V. Kulakovskiy, D.V. Kuprash; Multiple single nucleotide polymorphisms in the first intron of the IL2RA gene affect transcription factor binding and enhancer activity. *Gene*, 602:50-56, doi: 10.1016/j.gene.2016.11.032
2. (2017) M.A. Afanasyeva, L.V. Putlyaeva, D.E. Demin, I.V. Kulakovskiy, I.E. Vorontsov, M.V. Fridman, V.J. Makeev, D.V. Kuprash, A.M. Schwartz; The single nucleotide variant rs12722489 determines differential estrogen receptor binding and enhancer properties of an IL2RA intronic region. *PloS One*, 12(2): e0172681, doi:10.1371/journal.pone.0172681
3. (2016) I.V. Kulakovskiy, I.E. Vorontsov, I.S. Yevshin, A.V. Soboleva, A.S. Kasianov, H. Ashoor, W. Ba-Alawi, V.B. Bajic, Y.A. Medvedeva, F.A. Kolpakov, V.J. Makeev; HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.*, 44(D1):D116-25, doi: 10.1093/nar/gkv1249
4. (2016) I.E. Vorontsov, G.Khimulya, E.N. Lukianova, D.D. Nikolaeva, I.A. Eliseeva, I.V. Kulakovskiy, V.J. Makeev; Negative selection maintains transcription factor binding motifs in human cancer. *BMC Genomics*, 17(S.2):395, doi: 10.1186/s12864-016-2728-9
5. (2016) N. Zolotarev, A. Fedotova, O. Kyrchanova, A. Bonchuk, A.A. Penin, A.S. Lando, I.A. Eliseeva, I.V. Kulakovskiy, O. Maksimenko, P. Georgiev; Architectural proteins Pita, Zw5, and ZIPIC contain homodimerization domain and support specific long-range interactions in Drosophila. *Nucleic Acids Res.*, doi: 10.1093/nar/gkw371
6. (2016) A.M. Schwartz, L.V. Putlyaeva, M.Covich, A.V. Klepikova, K.A. Akulich, I.E. Vorontsov, K.V. Korneev, S.E. Dmitriev, O.L. Polanovsky, S.P. Sidorenko, I.V. Kulakovskiy, D.V. Kuprash; Early B-cell factor 1 (EBF1) is critical for transcriptional control of SLAMF1 gene in human B cells. *BBA-Gene Regulatory Mechanisms*, doi: 10.1016/j.bbagrm.2016.07.004
7. (2015) K. Kozlov, V.V. Gursky, I.V. Kulakovskiy, A. Dymova, M. Samsonova; Analysis of functional importance of binding sites in the Drosophila gap gene network model. *BMC Genomics*, 16 Suppl 13:S7. doi: 10.1186/1471-2164-16-S13-S7
8. (2015) D. Papatsenko, H. Darr, I.V. Kulakovskiy, A. Waghray, V.J. Makeev, B.D. MacArthur, I.R. Lemischka; Single-Cell Analyses of ESCs Reveal Alternative Pluripotent Cell States and Molecular Mechanisms that Control Self-Renewal. *Stem Cell Reports*, 5(2):207-20. doi: 10.1016/j.stemcr.2015.07.004
9. (2015) Y.A. Medvedeva, A. Lennartsson, R. Ehsani, I.V. Kulakovskiy, I.E. Vorontsov, P. Panahandeh, G. Khimulya, T. Kasukawa, The FANTOM Consortium and F. Drabløs; EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database*, bav067, doi:10.1093/database/bav067

10. (2014) K. Kozlov, V. Gursky, I. Kulakovskiy, M. Samsonova; Sequence-based model of gap gene regulatory network. *BMC Genomics*, 15(S.12):S6, doi:10.1186/1471-2164-15-S12-S6
11. (2014) A.R.R. Forrest, H. Kawaji, M. Rehli, J.K. Baillie, M.J.L. de Hoon, V. Haberle, T. Lassmann, I.V. Kulakovskiy, M. Lizio, M. Itoh *et al.*; A promoter-level mammalian expression atlas. *Nature*, 507: 462–470, doi: 10.1038/nature13182
12. (2014) Y.A. Medvedeva, A.M. Khamis, I.V. Kulakovskiy, W. Ba-Alawi, M.S.I. Bhuyan, H. Kawaji, T. Lassmann, M. Harbers, A.R.R. Forrest, V.B. Bajic, The FANTOM consortium; Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics*, 15(1): 119, doi: 10.1186/1471-2164-15-119
13. (2014) V.G. Levitsky, I.V. Kulakovskiy, N.I. Ershov, D.Y. Oschepkov, V.J. Makeev, T.C. Hodgman, T.I. Merkulova; Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genomics*, 15(1): 80, doi: 10.1186/1471-2164-15-80
14. (2013) I.A. Eliseeva, I.E. Vorontsov, K.E. Babeyev, S.M. Buyanova, M.A. Sysoeva, F.A. Kondrashov, I.V. Kulakovskiy; In silico motif analysis suggests an interplay of transcriptional and translational control in mTOR response. *Translation*, 1(2), doi: 10.4161/trla.27469
15. (2013) I.E. Vorontsov, I.V. Kulakovskiy, V.J. Makeev; Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms for Molecular Biology*, 8(23), doi: 10.1186/1748-7188-8-23
16. (2013) I. Kulakovskiy, V. Levitsky, D. Oshchepkov, L. Bryzgalov, I. Vorontsov, V. Makeev; From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *Journal of Bioinformatics and Computational Biology*, 11(1): 1340004, doi: 10.1142/S0219720013400040
17. (2013) I.V. Kulakovskiy, Y.A. Medvedeva, U. Schaefer, A.S. Kasianov, I.E. Vorontsov, V.B. Bajic, V.J. Makeev; HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.*, 41(D1): D195-D202, doi: 10.1093/nar/gks1089
18. (2012) M. Ridinger-Saison, V. Boeva, P. Rimmele, I. Kulakovskiy, I. Gallais, B. Levavasseur, C. Paccard, P. Legoix-Ne, F. Morle, A. Nicolas, P. Hupe, E. Barillot, F. Moreau-Gachelin, C. Guillouf; Spi-1/PU.1 activates transcription through clustered DNA occupancy in erythroleukemia. *Nucleic Acids Res.*, 40(18): 8927-8941, doi: 10.1093/nar/gks659
19. (2011) I.V. Kulakovskiy, A.A. Belostotsky, A.S. Kasianov, N.G. Esipova, Y.A. Medvedeva, I.A. Eliseeva, V.J. Makeev; A deeper look into transcription regulatory code by preferred pair distance templates for transcription factor binding sites. *Bioinformatics*, 27(19): 2621-4, doi: 10.1093/bioinformatics/btr453
20. (2010) I.V. Kulakovskiy, V.A. Boeva, A.V. Favorov, V.J. Makeev; Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, 26(20): 2622-3, doi: 10.1093/bioinformatics/btq488

21. (2010) Y.A. Medvedeva, M.V. Fridman, N.J. Oparina, D.B. Malko, E.O. Ermakova, I.V. Kulakovskiy, A. Heinzel, V.J. Makeev; Intergenic, gene terminal, and intragenic CpG islands in the human genome. *BMC Genomics*, 11:48, doi: 10.1186/1471-2164-11-48

8.2. Статьи в рецензируемых российских журналах

вх. в список ВАК

22. (2011) И.В. Кулаковский, А.С. Касьянов, А.А. Белостоцкий, И.А. Елисеева, В.Ю. Макеев; Предпочтительные расстояния между участками ДНК, связывающими белковые факторы, регулирующие инициацию транскрипции. *Биофизика*, 56(1): 136-9

23. (2010) Ю.А. Медведева, И.В. Кулаковский, Н.Ю. Опарина, А.В. Фаворов, В.Ю. Макеев; Ассиметрия GC-состава в окрестностях стартов транскрипции (с участием ДНК-зависимой РНК-полимеразы II) и ее связь с расположением участков адсорбции белка Sp1 на ДНК. *Биофизика*, 55(6): 976-85

8.3. Приглашенные главы в книгах и сериях обзоров

24. (2014) I.V. Kulakovskiy, V.J. Makeev; Motif Discovery and Motif Finding in ChIP-Seq Data. Invited chapter in *Genome Analysis: Current Procedures and Applications*, edited by Maria Poptsova, Caister Academic Press, 83-100. ISBN: 978-1-908230-29-4

25. (2013) I.V. Kulakovskiy, V.J. Makeev; DNA Sequence Motif: A Jack of All Trades for ChIP-Seq Data. Invited chapter in *Advances in Protein Chemistry and Structural Biology*, Vol. 91, edited by Rossen Donev, Burlington: Academic Press, Elsevier Inc, 135-171. ISBN: 978-0-12-411637-5

8.4. Статьи в рецензируемых сборниках

26. (2015) I.E. Vorontsov, I.V. Kulakovskiy, G. Khimulya, D.D. Nikolaeva, V.J. Makeev, PERFECTOS-APE - Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation. In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOSTEC 2015), 102-108. ISBN 978-989-758-070-3, doi: 10.5220/0005189301020108

27. (2013) I.V. Kulakovskiy, V.G. Levitsky, D.G. Oschepkov, I.E. Vorontsov, V.J. Makeev, Learning Advanced TFBS Models from Chip-Seq Data - diChIPMunk: Effective Construction of Dinucleotide Positional Weight Matrices. In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOSTEC 2013), 146-150. ISBN 978-989-8565-35-8, doi: 10.5220/0004238201460150

8.5. Авторские доклады на конференциях

8.5.1. Пленарные и приглашенные доклады

28. (2017) 17-21 апреля, 21-я международная пущинская школа-конференция молодых учёных «Биология – наука 21 века», Пущино, Россия; Существует ли аналог генетического кода для некодирующих последовательностей? Анализ паттернов в регуляторных районах генома.
29. (2015) June 2-6, *IV International Conference "Modern Problems of Genetics, Radiobiology, Radioecology, and Evolution"* dedicated to the 115th anniversary of the birth of N.W.Timofeeff-Ressovsky and his international scientific school, St.Petersburg, Russia; Is there a nucleotide-level genetic code for gene expression control?
30. (2012) November 22-24, *The Third International Scientific and Practical Conference "Postgenomic methods of analysis in biology, and laboratory and clinical medicine"*, POSTGENOME'2012, Kazan, Russia; How transcription start site can affect translational regulation? A case study on mRNA targets of mTOR cascade.
31. (2012) June 25–29, *The Eighth International Conference on Bioinformatics of Genome Regulation and Structure/Systems Biology*, BGRS\SB'12, Novosibirsk, Russia; Comprehensive collection of human transcription factor binding site models.
32. (2010) November 8-10, *LIX Bioinformatics Colloquium*, Paris, France; Deep and wide digging for binding motifs in ChIP-Seq data.

8.5.2. Устные доклады

33. (2016) August 29-September 2, *The 10th International Conference on Bioinformatics of Genome Regulation and Structure/Systems Biology*, BGRS\SB'2016, Novosibirsk, Russia; I.E. Vorontsov, Y.A. Medvedeva, V.J. Makeev, I.V. Kulakovskiy; HOCOMOCO comprehensive model collection as a practical gateway to regulatory motif-ome of human and mouse transcription factors.
34. (2016) June 14-16, *SocBiN Bioinformatics 2016*, Moscow, Russia; I.V. Kulakovskiy; HOCOMOCO collection of transcription factor binding sites models: Expansion, enhancement and practical applications.
35. (2015) July 16-19, *Moscow Conference on Computational Molecular Biology* (MCCMB'15); I.E. Vorontsov, I.V. Kulakovskiy, G.N. Khimulya, D.D. Nikolaeva, V.J. Makeev; Selection pressure on breast cancer somatic mutations revealed by bioinformatics sequence analysis.
36. (2015) July 10-14, *23rd Annual International Conference on Intelligent Systems for Molecular Biology / 14th European Conference on Computational Biology*, ISMB/ECCB 2015, [VarI Special Interest Group], Dublin, Ireland; I.E. Vorontsov, I.V. Kulakovskiy, G. Khimulya, D. Nikolaeva, V.J. Makeev; Sequence analysis of regulatory variants reveals selection pressure on somatic mutations in breast cancer.
37. (2013) July 21-23, *21st Annual International Conference on Intelligent Systems for Molecular Biology / 12th European Conference on Computational Biology*, ISMB/ECCB 2013, [Regulatory Genomics Special Interest Group], Berlin, Germany;

I.V. Kulakovskiy; diChIPMunk: utilizing ChIP-Seq data to construct advanced dinucleotide models of transcription factor binding sites.

38. (2013) July 25-28, *Moscow Conference on Computational Molecular Biology*, MCCMB'13, Moscow, Russia; I.V. Kulakovskiy, V.G. Levitsky, D.G. Oshchepkov, I.E. Vorontsov, V.J. Makeev; From ChIP-Seq data to improved transcription factor binding sites models.
39. (2012) October 29 - November 2, FANTOM5 Timecourses meeting, Workshop: Motifs and composite elements for time-courses, RIKEN Yokohama Institute, Tsurumi, Japan; I.A. Eliseeva, K.E. Babeev, S.M. Buyanova, M.A. Sysoeva, F.A. Kondrashov, I.V. Kulakovskiy; Sequence analysis of TSSs for mRNAs translationally regulated by mTOR cascade.
40. (2011) July 15-19, *19th Annual International Conference on Intelligent Systems for Molecular Biology / 10th European Conference on Computational Biology*, [Bioinformatics for Regulatory Genomics Special Interest Group], Vienna, Austria; I.V. Kulakovskiy, A.A. Belostotsky, A.S. Kasianov, I.A. Eliseeva, V.J. Makeev; Preferred Pair Distance Templates reveal functional transcription factor binding sites.
41. (2011) June 14-18, *Albany 2011: The 17th conversation*, Albany, NY, USA; I.V. Kulakovskiy, A.A. Belostotsky, V.J. Makeev; Preferred Pair Distance Templates for Identification of Functional Binding Sites for Interacting Transcription Factors. *Jounal of Biomolecular Structure & Dynamics*, Volume 28, Issue #6, p.1127-8.

8.5.3. Стендовые доклады

42. (2014) March 27-28, *Integrative and Computational Biology / V IMPPC Annual Conference and 4DCellFate Workshop*, Barcelona, Spain; I.A. Eliseeva, I.E. Vorontsov, K.E. Babeyev, S.M. Buyanova, M.A. Sysoeva, F.A. Kondrashov, I.V. Kulakovskiy; mTOR mRNA targets and transcription-translation rendezvous: a peep through sequence analysis keyhole.
43. (2013) July 21-23, *21st Annual International Conference on Intelligent Systems for Molecular Biology / 12th European Conference on Computational Biology, ISMB/ECCB 2013*, Berlin, Germany; I.V. Kulakovskiy, V.G. Levitsky, D.G. Oshchepkov, I.E. Vorontsov, V.J. Makeev; diChIPMunk: deriving dinucleotide TFBS models from ChIP-Seq data.
44. (2012) September 9-12, *11th European Conference on Computational Biology*, ECCB'12, Basel, Switzerland; I.V. Kulakovskiy, Y.A. Medvedeva, A.S. Kasianov, I.E. Vorontsov, U. Schaefer, V.B. Bajic, V.J. Makeev; Comprehensive collection of human transcription factor binding sites models (HOCOMOCO).
45. (2011) July 21-24, *The International Moscow Conference on Computational Molecular Biology*, MCCMB'11, Moscow, Russia; I.V. Kulakovskiy, A.A. Belostotsky, A.S. Kasianov, Y.A. Medvedeva, I.A. Eliseeva, V.J. Makeev; Preferred Pair Distance Templates for Analysis of Transcription Regulation Code, p.182-4
46. (2011) 11 мая, *Официальное открытие центра коллективного пользования (ЦКП) «Биоинформатика»*, Новосибирск, Академгородок, Россия; И.В. Кулаковский, А.С. Касьянов, В.Ю. Макеев; Алгоритмические проблемы анализа данных ChIP-Seq.

47. (2010) September 26-29, *9th European Conference on Computational Biology,* ECCB'10, Ghent, Belgium; I.V. Kulakovskiy, V.A. Boeva, A.V. Favorov, V.J. Makeev; Fast and accurate digging for binding motifs in ChIP-Seq data using ChIPMunk software.
48. (2010) June 20-27, *The 7th International Conference on Bioinformatics of Genome Regulation and Structure\Systems Biology*, BGRS\SB'10, Novosibirsk, Russia; I.V. Kulakovskiy, V.A. Boeva, A.V. Favorov, V.J. Makeev; ChIPMunk: discovery of transcription factor binding motifs in ChIP-Seq data. p.152
49. (2010) April 14-17, *The 4th ESF Conference on Functional Genomics & Disease*, Dresden, Germany; I.V. Kulakovskiy, V.A. Boeva, A.V. Favorov, V.J. Makeev; Motif discovery for transcription factor binding sites using a priori information on potential similarity regions at different resolution scales. *New Biotechnology*, Volume 27, Supplement 1, P1.52, S41

Благодарности

Эта работа была бы невозможной без поддержки, которую оказывали близкие родственники: Т.В. Кулаковская, И.А. Елисеева, Р.И. Кулаковский.

Исключительную роль сыграла всесторонняя поддержка начинаний и свобода научного поиска, предоставленная в лабораториях под рук. В.Г. Туманяна (ИМБ РАН) и В.Ю. Макеева (ИОГен РАН). Лично В.Ю. Макееву и В.Г. Туманяну выражаю глубокую и искреннюю признательность.

Благодарю дирекцию ИМБ РАН и лично акад. РАН А.А. Макарова за многолетнюю поддержку и толератное отношение к особенностям биоинформатической работы.

Хочется поблагодарить Ф.А. Кондрашова, М.С. Гельфанда, Д.А. Папаценко, Д.В. Купраша, М.А. Ройтберга, М.Г. Самсонову, Н.Г. Есипову, А. Фаворова, А. Зиновьева, Ф. Колпакова, В.Боеву, А. Касьянова, М. Фридман, А. Анашкину, Д. Ощепкова, В. Левицкого, И. Гроссе, А. Келя, А.Р.Р. Форреста, Ж. ван Хельдена, и многих других коллег, обсуждавших представленную работу на различных ее этапах, делившихся идеями, принимавших участие в наших проектах и приглашавших к участию в своих.

Отдельной строкой выражаю благодарность Ю.А. Медведевой и И. Воронцову, которые стартовали и довели до завершения ряд проектов, представленных в диссертации.

Исследования, которые легли в основу этой диссертации, были поддержаны различными научными фондами (включая Российский фонд фундаментальных исследований и Российский научный фонд), но ту самую возможность продуктивно работать в России по завершению аспирантуры мне подарил ликвидированнный ныне Фонд «Династия». Искренне благодарю исполнительного директора Фонда – А.Пиотровскую, и основателя Фонда – Д.Б. Зимина.

Низкий поклон людям, которые в значительной степени повлияли на мое мировозрение и интерес к науке. Это учителя общеобразовательной школы №1 города Пущино, среди которых особое место в сердце навсегда заняли И.В. Попкова, Т.С. Сергеева, Л.И. Салосина, Т.О. Ашчян. Это учителя музыки – В.И. Ковалинская, Т.В. Брынских, Г.Ф. Долгина, и В.В. Ромашин – привившие и развившие усидчивость и упорство. Это преподаватели Московского Государственного Университета Леса – сотрудники профильных кафедр: А.В. Корольков, Л.В. Королькова, В.А. Шачнев, А.И. Рубинштейн, И.Е. Сигалов, Ю.А. Демьянов, А.А. Малашин. Это Лермонтова Рената Александровна, человек, не совместимый с халтурой и, по счастливому стечению обстоятельств, мой учитель и моя бабушка.

Искренне благодарю за поддержку и участие всех коллег и друзей, которых не удалось упомянуть поименно.

9. Список литературы

1. Медведева Ю.А. и др. Ассиметрия GC-состава в окрестностях стартов транскрипции (с участием ДНК-зависимой РНК-полимеразы II) и ее связь с расположением участков адсорбции белка Sp1 на ДНК // Биофизика. 2010. Т. 55. № 6. С. 976–985.

2. Меркулов В.М., Меркулова Т.И. Регуляторные транскрипционные факторы могут контролировать процесс транскрипции на стадии элонгации пре-мРНК // Вавиловский журнал генетики и селекции. 2014. Т. 18. № 2. С. 329–337.

3. Hertz G.Z., Stormo G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences // Bioinformatics. 1999. Т. 15. С. 563–577.

4. Eskin E., Pevzner P.A. Finding composite regulatory patterns in DNA sequences // Bioinformatics. 2002. Т. 18 Suppl 1. С. S354-63.

5. Bergman C.M., Carlson J.W., Celniker S.E. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster // Bioinformatics. 2005. Т. 21. С. 1747–1749.

6. Fratkin E. и др. MotifCut: regulatory motifs finding with maximum density subgraphs // Bioinformatics. 2006. Т. 22. С. e150-7.

7. Leung H.C., Chin F.Y. Finding motifs from all sequences with and without binding sites // Bioinformatics. 2006. Т. 22. С. 2217–2223.

8. Sinha S., Blanchette M., Tompa M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences // BMC Bioinformatics. 2004. Т. 5. С. 170.

9. Khoury G., Gruss P. Enhancer elements // Cell. 1983. Т. 33. С. 313–314.

10. Hertz G.Z., Hartzell G W 3rd, Stormo G.D. Identification of consensus patterns in unaligned DNA sequences known to be functionally related // Comput Appl Biosci. 1990. Т. 6. С. 81–92.

11. Lawrence C.E., Reilly A.A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences // Proteins Struct. Funct. Genet. 1990. Т. 7. № 1. С. 41–51.

12. Beckstette M. и др. Fast index based algorithms and software for matching position specific scoring matrices // BMC Bioinformatics. 2006. Т. 7. С. 389.

13. Mukherjee S. и др. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays // Nat Genet. 2004. Т. 36. № 12. С. 1331–1339.

14. Butler J.E.F., Kadonaga J.T. The RNA polymerase II core promoter: a key component in the regulation of gene expression // Genes Dev. 2002. Т. 16. № 20. С. 2583–2592.

15. Newburger D.E., Bulyk M.L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions // Nucleic Acids Res. 2009. Т. 37. № SUPPL. 1. С. D77-82.

16. Goodman S.D. и др. In vitro selection of integration host factor binding sites // J Bacteriol. 1999. Т. 181. С. 3246–3255.

17. Philippakis A.A. и др. Design of compact, universal DNA microarrays for protein binding microarray experiments // J Comput Biol. 2008. Т. 15. С. 655–665.

18. Fields D.S. и др. Quantitative specificity of the Mnt repressor // J Mol Biol. 1997. Т. 271. С. 178–194.

19. Helden J. van, Andre B., Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies // J Mol Biol. 1998. Т. 281. С. 827–842.

20. Vanet A. и др. Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori sigma(80) family of promoter signals // J Mol Biol. 2000. Т. 297. С. 335–353.

21. Hampshire A.J. и др. Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands // Methods. 2007. Т. 42. С. 128–140.

22. Stormo G.D. Consensus patterns in DNA // Methods Enzym. 1990. Т. 183. С. 211–221.

23. Thompson W. и др. Using the Gibbs Motif Sampler for phylogenetic footprinting // Methods Mol Biol. 2007. Т. 395. С. 403–424.

24. Oliphant A.R., Brandl C.J., Struhl K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein // Mol Cell Biol. 1989. Т. 9. С. 2944–2949.

25. Dai S.M. и др. Ligation-mediated PCR for quantitative in vivo footprinting // Nat Biotechnol. 2000. Т. 18. С. 1108–1111.

26. Berger M.F. и др. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities // Nat Biotechnol. 2006. Т. 24. С. 1429–1435.

27. Harr R., Haggstrom M., Gustafsson P. Search algorithm for pattern match analysis of nucleic acid sequences // Nucleic Acids Res. 1983. Т. 11. С. 2943–2957.

28. Mulligan M.E. и др. Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity // Nucleic Acids Res. 1984. Т. 12. С. 789–800.

29. Kolchanov N.A. и др. Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucleic Acids Res. 2002. Т. 30. С. 312–317.

30. Kel-Margoulis O. V и др. TRANSCompel: a database on composite regulatory elements in eukaryotic genes // Nucleic Acids Res. 2002. Т. 30. С. 332–334.

31. Matys V. и др. TRANSFAC: transcriptional regulation, from patterns to profiles // Nucleic Acids Res. 2003. Т. 31. С. 374–378.

32. Thompson W., Rouchka E.C., Lawrence C.E. Gibbs Recursive Sampler: finding transcription factor binding sites // Nucleic Acids Res. 2003. Т. 31. С. 3580–3585.

33. Sinha S., Tompa M. YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation // Nucleic Acids Res. 2003. Т. 31. С. 3586–3588.

34. Pavesi G. и др. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes // Nucleic Acids Res. 2004. Т. 32. С. W199-203.

35. Mueller P.R., Wold B. In vivo footprinting of a muscle specific enhancer by ligation mediated PCR // Science (80-). 1989. Т. 246. С. 780–786.

36. Blackwell T.K., Weintraub H. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection // Science (80-). 1990. Т. 250. С. 1104–1110.

37. Lawrence C.E. и др. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment // Science (80-). 1993. Т. 262. С. 208–214.

38. Abeel T. и др. GenomeView: A next-generation genome browser // Nucleic Acids Res. 2012. Т. 40. № 2.

39. Abramowitz M., Stegun I.A. Handbook of Mathematical Functions. New York: Dover Publications, 1972. 1006 с.

40. Afanasyeva M.A. и др. The single nucleotide variant rs12722489 determines differential estrogen receptor binding and enhancer properties of an IL2RA intronic region // PLoS One. 2017. Т. 12. № 2. С. e0172681.

41. Agius P. и др. High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions // PLoS Comput Biol. 2010. Т. 6. № 9.

42. Alamanova D., Stegmaier P., Kel A. Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. // BMC Bioinformatics. 2010. Т. 11. С. 225.

43. Alexandrov L.B. и др. Signatures of mutational processes in human cancer. // Nature. 2013. Т. 500. № 7463. С. 415–21.

44. Andersson R. и др. An atlas of active enhancers across human cell types and tissues // Nature. 2014. Т. 507. № 7493. С. 455–461.

45. Arnosti D.N., Kulkarni M.M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? // J Cell Biochem. 2005. Т. 94. № 5. С. 890–898.

46. Astanehe A. и др. MKNK1 is a YB-1 target gene responsible for imparting trastuzumab resistance and can be blocked by RSK inhibition. // Oncogene. 2012. Т. 31. № 41. С. 4434–4446.

47. Auerbach R.K. и др. Mapping accessible chromatin regions using Sono-Seq // Proc. Natl. Acad. Sci. 2009. Т. 106. № 35. С. 14926–14931.

48. Avni D., Biberman Y., Meyuhas O. The 5’ terminal oligopyrimidine tract confers translational control on TOP mRNAs in a cell type- and sequence context-dependent manner. // Nucleic Acids Res. 1997. Т. 25. № 5. С. 995–1001.

49. Bailey T.L. Discovering novel sequence motifs with MEME. // Curr Protoc Bioinformatics. 2002. Т. Chapter 2. С. Unit 2.4.

50. Bailey T.L. и др. MEME SUITE: tools for motif discovery and searching. // Nucleic Acids Res. 2009. Т. 37. № SUPPL. 2. С. W202--8.

51. Bailey T.L. и др. The value of position-specific priors in motif discovery using MEME. // BMC Bioinformatics. 2010. Т. 11. С. 179.

52. Bailey T.L. и др. The MEME Suite // Nucleic Acids Res. 2015. Т. 43. № W1. С. W39-49.

53. Bailey T.L., Elkan C. Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers // Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. , 1994. С. 28–36.

54. Bailey T.L., Machanick P. Inferring direct DNA binding from ChIP-seq. // Nucleic Acids Res. 2012.

55. Barozzi I. и др. Fish the ChIPs: a pipeline for automated genomic annotation of ChIP-Seq data. // Biol Direct. 2011. Т. 6. С. 51.

56. Barski A., Zhao K. Genomic location analysis by ChIP-seq // J Cell Biochem. 2009. Т. 107. № 1. С. 11–18.

57. Bell R.J.A. и др. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer // Science (80-). 2015. Т. 348. № 6238. С. 1036–1039.

58. Benjamini Y., Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing // J R Stat. Soc. 1995. Т. 57. № 1. С. 289–300.

59. Benos P. V, Bulyk M.L., Stormo G.D. Additivity in protein-DNA interactions: how good an approximation is it? // Nucleic Acids Res. 2002. Т. 30. № 20. С. 4442–4451.

60. Benzer S. FINE STRUCTURE OF A GENETIC REGION IN BACTERIOPHAGE. // Proc Natl Acad Sci U S A. 1955. Т. 41. № 6. С. 344–54.

61. Berg O.G. Statistical ensembles for sequence variability // J Mol Biol. 1987. Т. 193. № 4. С. 743–750.

62. Berg O.G., Hippel P.H. von. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. // J Mol Biol. 1987. Т. 193. № 4. С. 723–750.

63. Berg O.G., Hippel P.H. von. Selection of DNA binding sites by regulatory proteins. // Trends Biochem Sci. 1988. Т. 13. № 6. С. 207–211.

64. Berkum N.L. van и др. Hi-C: a method to study the three-dimensional architecture of genomes. // J Vis Exp. 2010. № 39.

65. Berman H.M. и др. The Protein Data Bank // Nucleic Acids Res. 2000. Т. 28. № 1. С. 235–242.

66. Bi Y. и др. Tree-based position weight matrix approach to model transcription factor binding site profiles. // PLoS One. 2011. Т. 6. № 9. С. e24210.

67. Biberman Y., Meyuhas O. Substitution of just five nucleotides at and around the transcription start site of rat beta-actin promoter is sufficient to render the resulting transcript a subject for translational control. // FEBS Lett. 1997. Т. 405. № 3. С. 333–336.

68. Bin G.C. Le и др. Oct4 is required for lineage priming in the developing inner cell mass of the mouse blastocyst // Development. 2014. Т. 141. № 5. С. 1001–1010.

69. Bird A.P. DNA methylation versus gene expression // J Embryol Exp Morphol. 1984. Т. 83 Suppl. С. 31–40.

70. Birney E. и др. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. // Nature. 2007. Т. 447. № 7146. С. 799–816.

71. Blick A.J. и др. The Capacity to Act in Trans Varies Among Drosophila Enhancers // Genetics. 2016. Т. 203. № 1. С. 203–218.

72. Boeva V. и др. Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules. // Algorithms Mol Biol. 2007. Т. 2. С. 13.

73. Boeva V. и др. De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. // Nucleic Acids Res. 2010. Т. 38. № 11. С. e126.

74. Boeva V. и др. Nebula--a web-server for advanced ChIP-seq data analysis. // Bioinformatics. 2012. Т. 28. № 19. С. 2517–2519.

75. Boeva V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in Eukaryotic cells // Front. Genet. 2016. Т. 7. № FEB.

76. Borgel J. и др. Targets and dynamics of promoter DNA methylation during early mouse development. // Nat Genet. 2010. Т. 42. № 12. С. 1093–100.

77. Box G.E.P., Draper N.R. Empirical Model-Building and Response Surfaces. : John Wiley & Sons, 1987.

78. Boyes J., Bird A. Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. // EMBO J. 1992. Т. 11. С. 327–333.

79. Boyle A.P. и др. High-Resolution Mapping and Characterization of Open Chromatin across the Genome // Cell. 2008. Т. 132. № 2. С. 311–322.

80. Brown P. и др. MEME-LaB: motif analysis in clusters. // Bioinformatics. 2013. Т. 29. № 13. С. 1696–7.

81. Buck M.J., Lieb J.D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. // Genomics. 2004. Т. 83. № 3. С. 349–360.

82. Buenrostro J.D. и др. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. // Nat Methods. 2013. Т. 10. № 12. С. 1213–8.

83. Bulger M., Groudine M. Functional and mechanistic diversity of distal transcription enhancers // Cell. 2011. Т. 144. № 3. С. 327–339.

84. Burgess D.J. Epigenetics: Asymmetric complexity of the histone code // Nat. Rev. Genet. 2012. Т. 13. № 11. С. 756–756.

85. Bushnell D. a и др. Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. // Science. 2004. Т. 303. № 5660. С. 983–988.

86. Carlson B. Next Generation Sequencing: The Next Iteration of Personalized Medicine: Next generation sequencing, along with expanding databases like The Cancer Genome Atlas, has the potential to aid rational drug discovery and streamline clinical trials. // Biotechnol. Healthc. 2012. Т. 9. № 2. С. 21–5.

87. Carroll T.S. и др. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data // Front. Genet. 2014. Т. 5. № APR.

88. Casey G. и др. Next generation sequencing and a new era of medicine. // Gut. 2013. Т. 62. № 6. С. 920–32.

89. Cavalli M. и др. Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. // Hum Genet. 2016. Т. 135. № 5. С. 485–97.

90. Cawley S. и др. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs // Cell. 2004. Т. 116. № 4. С. 499–509.

91. Chatterjee R., Vinson C. CpG methylation recruits sequence specific transcription factors essential for tissue specific gene expression. // Biochim Biophys Acta. 2012. Т. 1819. № 7. С. 763–70.

92. Chen R. и др. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases // Nat Biotechnol. 2016. Т. 34. № 5. С. 531–538.

93. Chen X. и др. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. // Cell. 2008. Т. 133. № 6. С. 1106–1117.

94. Chia N.-Y. и др. A genome-wide RNAi screen reveals determinants of human embryonic stem cell

identity // Nature. 2010. Т. 468. № 7321. С. 316–320.

95. Choy M.-K. и др. Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. // BMC Genomics. 2010. Т. 11. № 1. С. 519.

96. Christensen B.C. и др. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CPG island context // PLoS Genet. 2009. Т. 5. № 8.

97. Chung D. и др. Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data // PLoS Comput Biol. 2011. Т. 7. № 7. С. e1002111.

98. Chung D. и др. dPeak: High Resolution Identification of Transcription Factor Binding Sites from PET and SET ChIP-Seq Data // PLoS Comput Biol. 2013. Т. 9. № 10. С. e1003246.

99. Cliften P.F. и др. Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis // Genome Res. 2001. Т. 11. С. 1175–1186.

100. Cramer P., Bushnell D. a, Kornberg R.D. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. // Science (80-). 2001. Т. 292. № 5523. С. 1863–76.

101. D'haeseleer P. What are DNA sequence motifs? // Nat Biotechnol. 2006. Т. 24. № 4. С. 423–425.

102. Dabrowski M. и др. Optimally choosing PWM motif databases and sequence scanning approaches based on ChIP-seq data. // BMC Bioinformatics. 2015. Т. 16. С. 140.

103. Daily K. и др. MotifMap: integrative genome-wide maps of regulatory motif sites for model species // BMC Bioinformatics. 2011. Т. 12. № 1. С. 495.

104. Daniel B., Nagy G., Nagy L. The intriguing complexities of mammalian gene regulation: How to link enhancers to regulated genes. Are we there yet? // FEBS Lett. 2014. Т. 588. № 15. С. 2379–2391.

105. Das M.K., Dai H.-K. A survey of DNA motif finding algorithms. // BMC Bioinformatics. 2007. Т. 8 Suppl 7. С. S21.

106. Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves // Proc. 23rd Int. Conf. Mach. Learn. -- ICML'06. 2006. С. 233–240.

107. Dawson M.A., Kouzarides T. Cancer epigenetics: From mechanism to therapy // Cell. 2012. Т. 150. № 1. С. 12–27.

108. Day W.H., McMorris F.R. Critical comparison of consensus methods for molecular sequences. // Nucleic Acids Res. 1992. Т. 20. № 5. С. 1093–1099.

109. Deplancke B. и др. The Genetics of Transcription Factor DNA Binding Variation // Cell. 2016. Т. 166. № 3. С. 538–554.

110. Dolfini D., Mantovani R. YB-1 (YBX1) does not bind to Y/CCAAT boxes in vivo. // Oncogene. 2012.

111. Dori-Bachash M. и др. Widespread promoter-mediated coordination of transcription and mRNA degradation. // Genome Biol. 2012. Т. 13. № 12. С. R114.

112. Dronamraju K.R. J.B.S. Haldane (1892-1964): centennial appreciation of a polymath. // Am J Hum Genet. 1992. Т. 51. № 4. С. 885–9.

113. Drouin J. Minireview: pioneer transcription factors in cell fate specification. // Mol Endocrinol. 2014. Т. 28. № 7. С. 989–98.

114. Dunham I. и др. An integrated encyclopedia of DNA elements in the human genome. // Nature. 2012. Т. 489. № 7414. С. 57–74.

115. Eckhardt F. и др. DNA methylation profiling of human chromosomes 6, 20 and 22. // Nat Genet. 2006. Т. 38. № 12. С. 1378–85.

116. Ehrlich M. и др. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells // Nucleic Acids Res. 1982. Т. 10. № 8. С. 2709–2721.

117. Eicher J.D. и др. GRASP v2.0: An update on the Genome-Wide Repository of Associations between SNPs and Phenotypes // Nucleic Acids Res. 2015. Т. 43. № D1. С. D799–D804.

118. Eliseeva I. и др. In silico motif analysis suggests an interplay of transcriptional and translational control in mTOR response // Translation. 2013. Т. 1. № 2. С. 18–24.

119. Ettwiller L. и др. Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. // Nat Methods. 2007. Т. 4. № 7. С. 563–565.

120. Ewing B. и др. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. // Genome Res. 1998. Т. 8. № 3. С. 175–85.

121. Farh K.K.-H. и др. Genetic and epigenetic fine mapping of causal autoimmune disease variants. // Nature. 2015. Т. 518. № 7539. С. 337–43.

122. Farley E.K. и др. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. // Proc Natl Acad Sci U S A. 2016. Т. 113. № 23. С. 6508–13.

123. Farthing C.R. и др. Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes // PLoS Genet. 2008. Т. 4. № 6.

124. Favorov A. и др. Exploring massive, genome scale datasets with the genometricorr package // PLoS Comput Biol. 2012. Т. 8. № 5.

125. Favorov A. V. и др. A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. // Bioinformatics. 2005. Т. 21. № 10. С. 2240–2245.

126. Fawcett T. An introduction to ROC analysis // Pattern Recognit Lett. 2006. Т. 27. № 8. С. 861–874.

127. Fejes A.P. и др. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. // Bioinformatics. 2008. Т. 24. № 15. С. 1729–1730.

128. Feldman N. и др. G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. // Nat Cell Biol. 2006. Т. 8. № 2. С. 188–194.

129. Feng J. и др. Identifying ChIP-seq enrichment using MACS // Nat Protoc. 2012. Т. 7. № 9. С. 1728–1740.

130. Fernandez P.C. и др. Genomic targets of the human c-Myc protein // Genes Dev. 2003. Т. 17. № 9. С. 1115–1129.

131. Festuccia N. и др. The role of pluripotency gene regulatory network components in mediating transitions between pluripotent cell states // Curr. Opin. Genet. Dev. 2013. Т. 23. № 5. С. 504–511.

132. Flicek P., Birney E. Sense from sequence reads: methods for alignment and assembly. // Nat Methods. 2009. Т. 6. № 11 Suppl. С. S6--S12.

133. Forrest A.R.R. и др. A promoter-level mammalian expression atlas. // Nature. 2014. Т. 507. № 7493. С. 462–70.

134. Friedman J.R., Kaestner K.H. The Foxa family of transcription factors in development and metabolism // Cell. Mol. Life Sci. 2006. Т. 63. № 19–20. С. 2317–2328.

135. Frith M.C. и др. A code for transcription initiation in mammalian genomes. // Genome Res. 2008. Т. 18. № 1. С. 1–12.

136. Frith M.C., Li M.C., Weng Z. Cluster-Buster: Finding dense clusters of motifs in DNA sequences // Nucleic Acids Res. 2003. Т. 31. № 13. С. 3666–3668.

137. Fujita P.A. и др. The UCSC Genome Browser database: update 2011. // Nucleic Acids Res. 2011. Т. 39. № Database issue. С. D876--82.

138. Fürst R.W. и др. A differentially methylated single CpG-site is correlated with estrogen receptor alpha transcription // J Steroid Biochem Mol Biol. 2012. Т. 130. № 1–2. С. 96–104.

139. Gagliardi A. и др. A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal // EMBO J. 2013. Т. 32. № 16. С. 2231–2247.

140. Galas D.J., Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. // Nucleic Acids Res. 1978. Т. 5. № 9. С. 3157–3170.

141. Gebhard C. и др. General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells // Cancer Res. 2010. Т. 70. № 4. С. 1398–1407.

142. Gershenzon N.I., Stormo G.D., Ioshikhes I.P. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. // Nucleic Acids Res. 2005. Т. 33. № 7. С. 2290–2301.

143. Giaquinta E. и др. Motif matching using gapped patterns // Theor Comp Sci. 2014. Т. 548. № С. С.

1–13.

144. Giaquinta E., Grabowski S., Ukkonen E. Fast Matching of Transcription Factor Motifs Using Generalized Position Weight Matrix Models. // J Comput Biol. 2013. Т. 20. № 9. С. 1–10.

145. Giorgetti L. и др. Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs. // Mol Cell. 2010. Т. 37. № 3. С. 418–428.

146. Girgis H.Z., Ovcharenko I. Predicting tissue specific cis-regulatory modules in the human genome using pairs of co-occurring motifs. // BMC Bioinformatics. 2012. Т. 13. № 1. С. 25.

147. Glick Y. и др. Integrated microfluidic approach for quantitative high-throughput measurements of transcription factor binding affinities // Nucleic Acids Res. 2015. Т. 44. № 6.

148. Gnatt A.L. и др. Structural basis of transcription: An RNA polymerase II elongation complex at 3.3 angstrom resolution // Science (80-). 2001. Т. 292. № 5523. С. 1876–1882.

149. Goentoro L. и др. The Incoherent Feedforward Loop Can Provide Fold-Change Detection in Gene Regulation // Mol Cell. 2009. Т. 36. № 5. С. 894–899.

150. Göke J. и др. Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development // PLoS Comput Biol. 2011. Т. 7. № 12.

151. Goodwin S., McPherson J.D., McCombie W.R. Coming of age: ten years of next-generation sequencing technologies. // Nat Rev Genet. 2016. Т. 17. № 6. С. 333–51.

152. Gossett A.J., Lieb J.D. DNA Immunoprecipitation (DIP) for the Determination of DNA-Binding Specificity. // CSH Protoc. 2008. Т. 2008. С. pdb.prot4972.

153. Gotea V. и др. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. // Genome Res. 2010. Т. 20. № 5. С. 565–577.

154. Grant C.E., Bailey T.L., Noble W.S. FIMO: Scanning for occurrences of a given motif // Bioinformatics. 2011. Т. 27. № 7. С. 1017–1018.

155. Grau J. и др. A general approach for discriminative de novo motif discovery from high-throughput data // Nucleic Acids Res. 2013. Т. 41. № 21.

156. Grau J., Grosse I., Keilwagen J. PRROC: Computing and visualizing Precision-recall and receiver operating characteristic curves in R // Bioinformatics. 2015. Т. 31. № 15. С. 2595–2597.

157. Gray K.A. и др. Genenames.org: the HGNC resources in 2013. // Nucleic Acids Res. 2013. Т. 41. № Database issue. С. D545--52.

158. Gu H. и др. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. // Nat Protoc. 2011. Т. 6. № 4. С. 468–481.

159. Guertin M.J. и др. Accurate prediction of inducible transcription factor binding intensities in Vivo // PLoS Genet. 2012. Т. 8. № 3.

160. Guillon N. и др. The oncogenic EWS-FLI1 protein binds in vivo GGAA microsatellite sequences with potential transcriptional activation function. // PLoS One. 2009. Т. 4. № 3. С. e4932.

161. Guo Y., Mahony S., Gifford D.K. High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints // PLoS Comput Biol. 2012. Т. 8. № 8. С. e1002638.

162. Gupta S. и др. Quantifying similarity between motifs. // Genome Biol. 2007. Т. 8. № 2. С. R24.

163. Ha N., Polychronidou M., Lohmann I. COPS: Detecting Co-Occurrence and Spatial Arrangement of Transcription Factor Binding Motifs in Genome-Wide Datasets // PLoS One. 2012. Т. 7. № 12. С. e52055.

164. Habib N. и др. A novel Bayesian DNA motif comparison method for clustering and retrieval // PLoS Comput Biol. 2008. Т. 4. № 2.

165. Harbison C.T. и др. Transcriptional regulatory code of a eukaryotic genome. // Nature. 2004. Т. 431. № 7004. С. 99–104.

166. Harrow J. и др. GENCODE: The reference human genome annotation for the ENCODE project // Genome Res. 2012. Т. 22. № 9. С. 1760–1774.

167. He X. и др. GABPα Binding to Overlapping ETS and CRE DNA Motifs Is Enhanced by CREB1:

Custom DNA Microarrays. // G3 (Bethesda, Md.). 2015. Т. 5. № 9. С. 1909–18.

168. Hehl R., Bülow L. AthaMap Web Tools for the Analysis of Transcriptional and Posttranscriptional Regulation of Gene Expression in Arabidopsis thaliana // Methods in molecular biology (Clifton, N.J.). , 2014. С. 139–156.

169. Heinemeyer T. и др. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. // Nucleic Acids Res. 1998. Т. 26. № 1. С. 362–367.

170. Heinz S. и др. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. // Mol Cell. 2010. Т. 38. № 4. С. 576–89.

171. Hochheimer A., Tjian R. Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. // Genes Dev. 2003. Т. 17. № 11. С. 1309–20.

172. Hogeweg P. The roots of bioinformatics in theoretical biology. // PLoS Comput Biol. 2011. Т. 7. № 3. С. e1002021.

173. Holm S. A Simple Sequentially Rejective Multiple Test Procedure // Scand. J. Stat. 1979. Т. 6. № 2. С. 65–70.

174. Hon C.-C. и др. An atlas of human long non-coding RNAs with accurate 5’ ends. // Nature. 2017. Т. 543. № 7644. С. 199–204.

175. Horak C.E., Snyder M. ChIP-chip: a genomic approach for identifying transcription factor binding sites. // Methods Enzymol. 2002. Т. 350. С. 469–483.

176. Hornstein E. и др. The expression of poly(A)-binding protein gene is translationally regulated in a growth-dependent fashion through a 5’-terminal oligopyrimidine tract motif. // J Biol Chem. 1999. Т. 274. № 3. С. 1708–1714.

177. Hsieh A.C. и др. The translational landscape of mTOR signalling steers cancer initiation and metastasis. // Nature. 2012. Т. 485. № 7396. С. 55–61.

178. Hsieh C.-L. и др. Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. // Proc Natl Acad Sci U S A. 2014. Т. 111. № 20. С. 7319–24.

179. Hu M. и др. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. // Nucleic Acids Res. 2010. Т. 38. № 7. С. 2154–2167.

180. Huang F.W. и др. Highly recurrent TERT promoter mutations in human melanoma. // Science. 2013. Т. 339. № 6122. С. 957–9.

181. Hume M.A. и др. UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions // Nucleic Acids Res. 2015. Т. 43. № D1. С. D117–D122.

182. Hunt L.T. Margaret O. Dayhoff // DNA. 1983. Т. 2. № 2. С. 97–98.

183. Hurst L.D. и др. A simple metric of promoter architecture robustly predicts expression breadth of human genes suggesting that most transcription factors are positive regulators // Genome Biol. 2014. Т. 15. № 7. С. 413.

184. Inoue F., Ahituv N. Decoding enhancers using massively parallel reporter assays // Genomics. 2015. Т. 106. № 3. С. 159–164.

185. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. // Nature. 2004. Т. 431. № 7011. С. 931–945.

186. Isakova A. и др. SMiLE-seq identifies binding motifs of single and dimeric transcription factors. // Nat Methods. 2017.

187. Ise T. и др. Transcription factor Y-box binding protein 1 binds preferentially to cisplatin-modified DNA and interacts with proliferating cell nuclear antigen. // Cancer Res. 1999. Т. 59. № 2. С. 342–6.

188. Istrail S., De-Leon S.B.-T., Davidson E.H. The regulatory genome and the computer. // Dev Biol. 2007. Т. 310. № 2. С. 187–95.

189. Ivanova N. и др. Dissecting self-renewal in stem cells with RNA interference // Nature. 2006. Т. 442. № 7102. С. 533–538.

190. Iwasaki H. и др. Distinctive and indispensable roles of PU.1 in maintenance of hematopoietic stem cells and their differentiation // Blood. 2005. Т. 106. № 5. С. 1590–1600.

191. Jacob F., Monod J. Genetic regulatory mechanisms in the synthesis of proteins. // J Mol Biol. 1961. Т. 3. С. 318–56.

192. Jacobson E.M. и др. Structure of Pit-1 POU domain bound to DNA as a dimer: Unexpected arrangement and flexibility // Genes Dev. 1997. Т. 11. № 2. С. 198–212.

193. Jaenisch R., Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. // Nat Genet. 2003. Т. 33 Suppl. № march. С. 245–254.

194. Jain D. и др. Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments // Nucleic Acids Res. 2015. Т. 43. № 14. С. 6959–6968.

195. Jankowski A., Tiuryn J., Prabhakar S. Romulus: Robust multi-state identification of transcription factor binding sites from DNase-seq data // Bioinformatics. 2016. С. btw209-.

196. Jefferies H.B., Thomas G. Elongation factor-1 alpha mRNA is selectively translated following mitogenic stimulation. // J Biol Chem. 1994. Т. 269. № 6. С. 4367–4372.

197. Jenks S. Navigating the genome's «dark matter». // J Natl Cancer Inst. 2013. Т. 105. № 10. С. 673–4.

198. Jenuwein T., Allis C.D. Translating the histone code. // Science. 2001. Т. 293. № 5532. С. 1074–80.

199. Ji H. и др. An integrated software system for analyzing ChIP-chip and ChIP-seq data. // Nat Biotechnol. 2008. Т. 26. № 11. С. 1293–1300.

200. Jiang P. и др. Inference of transcriptional regulation in cancers. // Proc Natl Acad Sci U S A. 2015. Т. 112. № 25. С. 7731–6.

201. Johnson D.S. и др. Genome-wide mapping of in vivo protein-DNA interactions. // Science. 2007. Т. 316. № 5830. С. 1497–1502.

202. Johnson S.C., Rabinovitch P.S., Kaeberlein M. mTOR is a key modulator of ageing and age-related disease. // Nature. 2013. Т. 493. № 7432. С. 338–345.

203. Jolma A. и др. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. // Genome Res. 2010. Т. 20. № 6. С. 861–873.

204. Jolma A. и др. DNA-binding specificities of human transcription factors. // Cell. 2013. Т. 152. № 1–2. С. 327–339.

205. Jones P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. // Nat Rev Genet. 2012. Т. 13. № 7. С. 484–92.

206. Jones P.A., Takai D. The role of DNA methylation in mammalian epigenetics // Science (80-). 2001. Т. 293. № 0036–8075 (Print). С. 1068–1070.

207. Jothi R. и др. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. // Nucleic Acids Res. 2008. Т. 36. № 16. С. 5221–5231.

208. Junion G. и др. A transcription factor collective defines cardiac cell fate and reflects lineage history // Cell. 2012. Т. 148. № 3. С. 473–486.

209. Kaestner K.H. The FoxA factors in organogenesis and differentiation // Curr. Opin. Genet. Dev. 2010. Т. 20. № 5. С. 527–532.

210. Kähärä J., Lähdesmäki H. Evaluating a linear k-mer model for protein-DNA interactions using high-throughput SELEX data. // BMC Bioinformatics. 2013. С. S2.

211. Kanamori-Katayama M. и др. Unamplified cap analysis of gene expression on a single-molecule sequencer. // Genome Res. 2011. Т. 21. № 7. С. 1150–1159.

212. Kankainen M., Löytynoja A. MATLIGN: a motif clustering, comparison and matching tool. // BMC Bioinformatics. 2007. Т. 8. С. 189.

213. Kaplan T. и др. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. // PLoS Genet. 2011. Т. 7. № 2. С. e1001290.

214. Kapranov P. и др. RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription // Science (80-). 2007. Т. 316. № 5830. С. 1484–1488.

215. Kato M. и др. Identifying combinatorial regulation of transcription factors and binding motifs // Genome Biol. 2004. Т. 5. № 8. С. R56.

216. Kawaji H. и др. Update of the FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. // Nucleic Acids Res. 2011. Т. 39. № Database issue. С. D856--60.

217. Keilwagen J. и др. MotifAdjuster: a tool for computational reassessment of transcription factor binding site annotations // Genome Biol. 2009. Т. 10. № 5. С. R46.

218. Keilwagen J., Grau J. Varying levels of complexity in transcription factor binding motifs. // Nucleic Acids Res. 2015.

219. Kharchenko P. V, Tolstorukov M.Y., Park P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. // Nat Biotechnol. 2008. Т. 26. № 12. С. 1351–1359.

220. Kheradpour P. и др. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay // Genome Res. 2013. Т. 23. № 5. С. 800–811.

221. Kheradpour P., Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments // Nucleic Acids Res. 2014. Т. 42. № 5. С. 2976–2987.

222. Khurana E. и др. Integrative annotation of variants from 1092 humans: application to cancer genomics. // Science. 2013. Т. 342. № 6154. С. 1235587.

223. Kibet C.K., Machanick P. Transcription factor motif quality assessment requires systematic comparative analysis // F1000Research. 2015. Т. 4.

224. Killela P.J. и др. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. // Proc Natl Acad Sci U S A. 2013. Т. 110. № 15. С. 6021–6.

225. Kim H. и др. A short survey of computational analysis methods in analysing ChIP-seq data. // Hum Genomics. 2011. Т. 5. № 2. С. 117–123.

226. Kim J.L., Nikolov D.B., Burley S.K. Co-crystal structure of TBP recognizing the minor groove of a TATA element // Nature. 1993. Т. 365. № 6446. С. 520–527.

227. Kim T.K., Shiekhattar R. Architectural and Functional Commonalities between Enhancers and Promoters // Cell. 2015. Т. 162. № 5. С. 948–959.

228. Kim Y. и др. Crystal structure of a yeast TBP/TATA-box complex. // Nature. 1993. Т. 365. № 6446. С. 512–20.

229. Kirsanov D.D. и др. NPIDB: Nucleic acid-Protein Interaction DataBase. // Nucleic Acids Res. 2013. Т. 41. № Database issue. С. D517-23.

230. Kleene K.C. и др. Alternative patterns of transcription and translation of the ribosomal protein L32 mRNA in somatic and spermatogenic cells in mice. // Exp Cell Res. 2003. Т. 291. № 1. С. 101–110.

231. Kok W., Oon Y.B., Lee N.K. Sequence logo visualization based on Gestalt perception (Novices vs experts) // 3rd International Conference of the South East Asian Network of Ergonomics Societies, SEANES2014. , 2014.

232. Korhonen J. и др. MOODS: fast search for position weight matrix matches in DNA sequences. // Bioinformatics. 2009. Т. 25. № 23. С. 3181–2.

233. Korhonen J.H. и др. Fast motif matching revisited: high-order PWMs, SNPs and indels. // Bioinformatics. 2016. С. btw683.

234. Kozlov K. и др. Sequence-based model of gap gene regulatory network. // BMC Genomics. 2014. Т. 15 Suppl 1. № Suppl 12. С. S6.

235. Kozlov K. и др. Analysis of functional importance of binding sites in the Drosophila gap gene network model. // BMC Genomics. 2015. Т. 16 Suppl 1. С. S7.

236. Kramer A.D.I., Guillory J.E., Hancock J.T. Experimental evidence of massive-scale emotional contagion through social networks. // Proc Natl Acad Sci U S A. 2014. Т. 111. № 24. С. 8788–90.

237. Krebs W. и др. Optimization of transcription factor binding map accuracy utilizing knockout-mouse models // Nucleic Acids Res. 2014. Т. 42. № 21. С. 13051–13060.

238. Kulakovskiy I. V и др. Deep and wide digging for binding motifs in ChIP-Seq data. // Bioinformatics. 2010. Т. 26. № 20. С. 2622–2623.

239. Kulakovskiy I. V и др. A deeper look into transcription regulatory code by preferred pair distance templates for transcription factor binding sites. // Bioinformatics. 2011. Т. 27. № 19. С. 2621–2624.

240. Kulakovskiy I. V и др. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. // Nucleic Acids Res. 2013a. Т. 41. № Database issue. С. D195–D202.

241. Kulakovskiy I. V и др. From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites // J Bioinform Comput Biol. 2013b. Т. 11. № 1. С. 1340004.

242. Kulakovskiy I. V и др. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. // Nucleic Acids Res. 2016. Т. 44. № D1. С. D116-25.

243. Kulakovskiy I. V., Favorov A. V., Makeev V.J. Motif discovery and motif finding from genome-mapped DNase footprint data // Bioinformatics. 2009. Т. 25. № 18. С. 2318–2325.

244. Kulakovskiy I. V., Makeev V.J. DNA sequence motif: a jack of all trades for ChIP-Seq data. // Adv Protein Chem Struct Biol. 2013. Т. 91. С. 135–171.

245. Kulakovskiy I. V, Makeev V.J. Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources // Biophysics (Oxf). 2009. Т. 54. № 6. С. 667–674.

246. Kulakovskiy I.V. и др. Learning Advanced TFBS Models from Chip-Seq Data - diChIPMunk: Effective Construction of Dinucleotide Positional Weight Matrices. // Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOSTEC 2013). : SCITEPRESS, 2013c. С. 146–150.

247. Kulakovskiy I.V., Makeev V.J. Motif Discovery and Motif Finding in ChIP-Seq Data // Genome Analysis: Current Procedures and Applications / под ред. P. Maria. , 2014. С. 83–100.

248. Kullback S., Leibler R.A. On information and sufficiency // Ann. Math. Stat. 1951. Т. 22. № 1. С. 79–86.

249. Kuttippurathu L. и др. CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments. // Bioinformatics. 2011. Т. 27. № 5. С. 715–717.

250. Kwok J.B. Role of epigenetics in Alzheimer’s and Parkinson’s disease // Epigenomics. 2010. Т. 2. № 5. С. 671–682.

251. Lada A.G. и др. Genome-Wide Mutation Avalanches Induced in Diploid Yeast Cells by a Base Analog or an APOBEC Deaminase // PLoS Genet. 2013. Т. 9. № 9. С. e1003736.

252. Lai E. и др. Hepatocyte nuclear factor 3 alpha belongs to a gene family in mammals that is homologous to the Drosophila homeotic gene fork head // Genes Dev. 1991. Т. 5. № 3. С. 416–427.

253. Laird P.W. Principles and challenges of genome-wide DNA methylation analysis // Nat. Rev. Genet. 2010. Т. 11. № 3. С. 191–203.

254. Landa I. и др. Allelic variant at -79 (C-T) in CDKN1B (p27Kip1) confers an increased risk of thyroid cancer and alters mRNA levels. // Endocr Relat Cancer. 2010. Т. 17. № 2. С. 317–28.

255. Landt S.G. и др. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia // Genome Res. 2012. Т. 22. № 9. С. 1813–1831.

256. Lang M., Juan E. Binding site number variation and high-affinity binding consensus of Myb-SANT-like transcription factor Adf-1 in Drosophilidae. // Nucleic Acids Res. 2010. Т. 38. № 19. С. 6404–6417.

257. Langmead B. и др. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. // Genome Biol. 2009. Т. 10. № 3. С. R25.

258. Langmead B., Salzberg S.L. Fast gapped-read alignment with Bowtie 2. // Nat Methods. 2012. Т. 9. № 4. С. 357–9.

259. Larsson E., Lindahl P., Mostad P. HeliCis: a DNA motif discovery tool for colocalized motif pairs with periodic spacing. // BMC Bioinformatics. 2007. Т. 8. С. 418.

260. Lawrenson K. и др. Common Variants at the CHEK2 Gene Locus and Risk of Epithelial Ovarian Cancer. // Carcinogenesis. 2015.

261. Leleu M., Lefebvre G., Rougemont J. Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. // Brief Funct Genomics. 2010. Т. 9. № 5–6. С. 466–476.

262. Lelli K.M., Slattery M., Mann R.S. Disentangling the Many Layers of Eukaryotic Transcriptional Regulation. // Annu Rev Genet. 2012.

263. Lemetre C., Zhang Z.D. A brief introduction to tiling microarrays: principles, concepts, and applications. // Methods Mol Biol. 2013. Т. 1067. С. 3–19.

264. Lemon B., Tjian R. Orchestrated response: a symphony of transcription factors for gene control // Genes Dev. 2000. Т. 14. № 20. С. 2551–2569.

265. Lenhard B., Sandelin A., Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. // Nat Rev Genet. 2012. № March.

266. Lerdrup M. и др. An interactive environment for agile analysis and visualization of ChIP-sequencing data. // Nat Struct Mol Biol. 2016. Т. 23. № 4. С. 349–57.

267. Lettice L.A. и др. Opposing functions of the ETS factor family define Shh spatial expression in limb buds and underlie polydactyly. // Dev Cell. 2012. Т. 22. № 2. С. 459–467.

268. Levine M., Cattoglio C., Tjian R. Looping back to leap forward: Transcription enters a new era. , 2014.

269. Levitskii V. и др. Recognition of transcription factor binding sites by the SiteGA method // Biophysics (Oxf). 2006. Т. 51. № 4. С. 570, 565.

270. Levitsky V.G. и др. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. // BMC Bioinformatics. 2007. Т. 8. С. 481.

271. Levitsky V.G. и др. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. // BMC Genomics. 2014. Т. 15. № 1. С. 80.

272. Levitsky V.G. и др. Hidden heterogeneity of transcription factor binding sites: A case study of SF-1 // Comput Biol Chem. 2016. Т. 64. С. 19–32.

273. Levo M. и др. Unraveling determinants of transcription factor binding outside the core binding site // Genome Res. 2015. С. gr.185033.114.

274. Levo M., Segal E. In pursuit of design principles of regulatory sequences // Nat. Rev. Genet. 2014. Т. 15. № 7. С. 453–468.

275. Lewis E.B. Pseudoallelism and gene evolution. // Cold Spring Harb Symp Quant Biol. 1951. Т. 16. С. 159–74.

276. Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. // Bioinformatics. 2009. Т. 25. № 14. С. 1754–1760.

277. Li H., Homer N. A survey of sequence alignment algorithms for next-generation sequencing. // Brief Bioinform. 2010. Т. 11. № 5. С. 473–483.

278. Li X.-Y. и др. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. // Genome Biol. 2011a. Т. 12. № 4. С. R34.

279. Li Y. и др. MDM2 SNP309 is associated with endometrial cancer susceptibility: a meta-analysis. // Hum. cell. 2011b. Т. 24. № 2. С. 57–64.

280. Liao J.C. и др. Network component analysis: reconstruction of regulatory signals in biological systems. // Proc Natl Acad Sci U S A. 2003. Т. 100. № 26. С. 15522–7.

281. Lifanov A.P. и др. Homotypic regulatory clusters in Drosophila. // Genome Res. 2003. Т. 13. № 4. С. 579–588.

282. Lifton R.P. и др. The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. // Cold Spring Harb Symp Quant Biol. 1978. Т. 42 Pt 2. С. 1047–1051.

283. Lin Y. и др. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. // BMC Genomics. 2016. Т. 17. С. 28.

284. Linhart C., Halperin Y., Shamir R. Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. // Genome Res. 2008. Т. 18. № 7. С. 1180–1189.

285. Lipkus A.H. A proof of the triangle inequality for the Tanimoto distance // J. Math. Chem. 1999. Т. 26. № 1. С. 263–265.

286. Lister R. и др. Human DNA methylomes at base resolution show widespread epigenomic differences. // Nature. 2009. Т. 462. № 7271. С. 315–22.

287. Liu E.T., Pott S., Huss M. Q&A: ChIP-seq technologies and the study of gene regulation. // BMC Biol. 2010. Т. 8. № 1. С. 56.

288. Liu X. и др. DIP-chip: rapid and accurate determination of DNA-binding specificity. // Genome Res. 2005. Т. 15. № 3. С. 421–427.

289. Liu Z.-P. и др. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. // Database (Oxford). 2015. Т. 2015.

290. Loh Y.-H. и др. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. // Nat Genet. 2006. Т. 38. № 4. С. 431–440.

291. Loo K.M.J. van и др. Transcriptional regulation of T-type calcium channel CaV3.2: bi-directionality by early growth response 1 (Egr1) and repressor element 1 (RE-1) protein-silencing transcription factor (REST). // J Biol Chem. 2012. Т. 287. № 19. С. 15489–15501.

292. Louder R.K. и др. Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. // Nature. 2016. Т. 531. № 7596. С. 604–9.

293. Ludwig M.Z., Patel N.H., Kreitman M. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change // Development. 1998. Т. 125. С. 949–958.

294. Lun D.S. и др. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. // Genome Biol. 2009. Т. 10. № 12. С. R142.

295. Luscombe N.M. и др. An overview of the structures of protein-DNA complexes // Genome Biol. 2000. Т. 1. № 1. С. reviews001.1.

296. Luscombe N.M., Laskowski R. a, Thornton J.M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. // Nucleic Acids Res. 2001. Т. 29. № 13. С. 2860–2874.

297. Lyabin D.N., Eliseeva I.A., Ovchinnikov L.P. YB-1 Synthesis Is Regulated by mTOR Signaling Pathway. // PLoS One. 2012. Т. 7. № 12. С. e52527.

298. Ma X. и др. A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. // Nucleic Acids Res. 2012. Т. 40. № 7. С. e50.

299. Ma X. и др. Reliable scaling of position weight matrices for binding strength comparisons between transcription factors. // BMC Bioinformatics. 2015. Т. 16. С. 265.

300. Maaskola J., Rajewsky N. Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models // Nucleic Acids Res. 2014. Т. 42. № 21. С. 12995–13011.

301. Machanick P., Bailey T.L. MEME-ChIP: motif analysis of large DNA datasets. // Bioinformatics. 2011. Т. 27. № 12. С. 1696–1697.

302. Macintyre G. и др. is-rSNP: a novel technique for in silico regulatory SNP detection. // Bioinformatics. 2010. Т. 26. № 18. С. i524--30.

303. Macleod D. и др. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island // Genes Dev. 1994. Т. 8. № 19. С. 2282–2292.

304. MacQuarrie K.L. и др. Genome-wide transcription factor binding: beyond direct target regulation. // Trends Genet. 2011. Т. 27. № 4. С. 141–8.

305. Mahony S., Benos P. V. STAMP: A web tool for exploring DNA-binding motif similarities // Nucleic Acids Res. 2007. Т. 35. № SUPPL.2.

306. Maksimenko O. и др. Two new insulator proteins, Pita and ZIPIC, target CP190 to chromatin // Genome Res. 2015. Т. 25. № 1. С. 89–99.

307. Manke T., Heinig M., Vingron M. Quantifying the effect of sequence variation on regulatory interactions. // Hum Mutat. 2010. Т. 31. № 4. С. 477–83.

308. Marchand B., Bajic V.B., Kaushik D.K. Highly scalable ab initio genomic motif identification // 2011 Int. Conf. High Perform. Comput. Networking, Storage Anal. 2011. С. 1–10.

309. Marinov G.K. и др. Large-scale quality analysis of published ChIP-seq data. // G3 (Bethesda, Md.). 2014. Т. 4. № 2. С. 209–23.

310. Mathelier A. и др. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. // Nucleic Acids Res. 2014. Т. 42. № Database issue. С. D142-7.

311. Mathelier A. и др. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles // Nucleic Acids Res. 2015a. Т. 44. № November 2015. С. gkv1176.

312. Mathelier A. и др. Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. // Genome Biol. 2015b. Т. 16. № 1. С. 84.

313. Mathelier A., Wasserman W.W. The next generation of transcription factor binding site prediction. // PLoS Comput Biol. 2013. Т. 9. № 9. С. e1003214.

314. Mathis D.J., Chambon P. The SV40 early region TATA box is required for accurate in vitro initiation of transcription. // Nature. 1981. Т. 290. № 5804. С. 310–5.

315. Matys V. и др. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. // Nucleic Acids Res. 2006. Т. 34. № Database issue. С. D108--10.

316. McLean C.Y. и др. GREAT improves functional interpretation of cis-regulatory regions. // Nat Biotechnol. 2010. Т. 28. № 5. С. 495–501.

317. Medina-Rivera A. и др. Theoretical and empirical quality assessment of transcription factor-binding motifs // Nucleic Acids Res. 2011. Т. 39. № 3. С. 808–824.

318. Medina-Rivera A. и др. RSAT 2015: Regulatory Sequence Analysis Tools. // Nucleic Acids Res. 2015. Т. 43. № W1. С. W50-6.

319. Medvedeva Y.A. и др. Intergenic, gene terminal, and intragenic CpG islands in the human genome. // BMC Genomics. 2010. Т. 11. С. 48.

320. Medvedeva Y.A. и др. Effects of cytosine methylation on transcription factor binding sites. // BMC Genomics. 2014. Т. 15. С. 119.

321. Medvedeva Y.A. и др. EpiFactors: A comprehensive database of human epigenetic factors and complexes // Database. 2015. Т. 2015.

322. Megraw M. и др. A transcription factor affinity-based code for mammalian transcription initiation. // Genome Res. 2009. Т. 19. № 4. С. 644–56.

323. Meissner A. и др. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis // Nucleic Acids Res. 2005. Т. 33. № 18. С. 5868–5877.

324. Melton C. и др. Recurrent somatic mutations in regulatory regions of human cancer genomes // Nat Genet. 2015. Т. 47. № 7. С. 710–6.

325. Mercer T.R., Dinger M.E., Mattick J.S. Long non-coding RNAs: insights into functions. // Nat Rev Genet. 2009. Т. 10. С. 155–159.

326. Mercier E. и др. An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. // PLoS One. 2011. Т. 6. № 2. С. e16432.

327. Merika M., Thanos D. Enhanceosomes // Curr Opin Genet Dev. 2001. Т. 11. № 2. С. 205–208.

328. Meyer C.A., Liu X.S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology // Nat. Rev. Genet. 2014. Т. 15. № 11. С. 709–721.

329. Meyuhas O. Synthesis of the translational apparatus is regulated at the translational level. // Eur J Biochem. 2000. Т. 267. № 21. С. 6321–6330.

330. Mistri T.K. и др. Selective influence of Sox2 on POU transcription factor binding in embryonic and neural stem cells // EMBO Rep. 2015. Т. 16. № 9. С. 1177–1191.

331. Monteiro P.T. и др. YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae. // Nucleic Acids Res. 2008. Т. 36. № Database issue. С. D132--6.

332. Mora A. и др. In the loop: promoter-enhancer interactions and bioinformatics. // Brief Bioinform. 2015. № July. С. 1–16.

333. Moreau-Gachelin F. и др. Spi-1/PU.1 transgenic mice develop multistep erythroleukemias. // Mol Cell Biol. 1996. Т. 16. № 5. С. 2453–63.

334. Morozov A. V, Siggia E.D. Connecting protein structure with predictions of regulatory sites. // Proc Natl Acad Sci U S A. 2007. Т. 104. № 17. С. 7068–7073.

335. Morozov V.Y., Ioshikhes I.P. Optimized Position Weight Matrices in Prediction of Novel Putative Binding Sites for Transcription Factors in the Drosophila melanogaster Genome // PLoS One. 2013. Т. 8. № 8.

336. Morris Q., Bulyk M.L., Hughes T.R. Jury remains out on simple models of transcription factor

specificity. // Nat Biotechnol. 2011. Т. 29. № 6. С. 483–4.

337. Muerdter F., Stark A. Genomics: Hiding in plain sight // Nature. 2014. Т. 512. № 7515. С. 374–375.

338. Mukhopadhyay R. и др. The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide // Genome Res. 2004. Т. 14. № 8. С. 1594–1602.

339. Müller H.P., Schaffner W. Transcriptional enhancers can act in trans. // Trends Genet. 1990. Т. 6. № 9. С. 300–4.

340. Murakami K., Kojima T., Sakaki Y. Assessment of clusters of transcription factor binding sites in relationship to human promoter, CpG islands and gene expression. // BMC Genomics. 2004. Т. 5. № 1. С. 16.

341. Myatt S.S., Lam E.W.-F. The emerging roles of forkhead box (Fox) proteins in cancer. // Nat Rev Cancer. 2007. Т. 7. № 11. С. 847–59.

342. Myers E.W., Miller W. Approximate matching of regular expressions. // Bull Math Biol. 1989. Т. 51. № 1. С. 5–37.

343. Myers R.M. и др. A user's guide to the encyclopedia of DNA elements (ENCODE). // PLoS Biol. 2011. Т. 9. № 4. С. e1001046.

344. Nakato R., Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation // Brief Bioinform. 2016. С. bbw023.

345. Nan X. и др. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex // Nature. 1998. Т. 393. № 6683. С. 386–389.

346. Natoli G., Andrau J.-C. Noncoding Transcription at Enhancers: General Principles and Functional Models // Annu Rev Genet. 2012. Т. 46. № 1. С. 1–19.

347. Nechaev S., Adelman K. Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. // Biochim Biophys Acta. 2011. Т. 1809. № 1. С. 34–45.

348. Neph S. и др. An expansive human regulatory lexicon encoded in transcription factor footprints. // Nature. 2012. Т. 489. № 7414. С. 83–90.

349. Neuwald A.F., Liu J.S., Lawrence C.E. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. // Protein Sci. 1995. Т. 4. № 8. С. 1618–1632.

350. Ng P. и др. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. // Nucleic Acids Res. 2006. Т. 34. № 12. С. e84.

351. Ng S.-Y.Y. и др. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. // Mol Cell. 2013. Т. 51. № 3. С. 349–59.

352. Ni Y. и др. Simultaneous SNP identification and assessment of allele-specific bias from ChIP-seq data. // BMC Genet. 2012. Т. 13. № 1. С. 46.

353. Nikolov D.B., Burley S.K. RNA polymerase II transcription initiation: a structural view // Proc Natl Acad Sci U S A. 1997. Т. 94. С. 15–22.

354. Nishida K., Frith M.C., Nakai K. Pseudocounts for transcription factor binding sites // Nucleic Acids Res. 2009. Т. 37. № 3. С. 939–944.

355. Normanno D. и др. Probing the target search of DNA-binding proteins in mammalian cells using TetR as model searcher. // Nat Commun. 2015. Т. 6. С. 7357.

356. Nutiu R. и др. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. // Nat Biotechnol. 2011. Т. 29. № 7. С. 659–664.

357. O'Malley R.C. и др. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape // Cell. 2016. Т. 165. № 5. С. 1280–1292.

358. Oda M. и др. DNA methylation regulates long-range gene silencing of an X-linked homeobox gene cluster in a lineage-specific manner // Genes Dev. 2006. Т. 20. № 24. С. 3382–3394.

359. Odom D.T. и др. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. // Nat Genet. 2007. Т. 39. № 6. С. 730–2.

360. Orr N. и др. Genome-wide association study identifies a common variant in RAD51B associated with male breast cancer risk. // Nat Genet. 2012. Т. 44. № 11. С. 1182–4.

361. Oshchepkov D.Y. и др. SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. // Nucleic Acids Res. 2004. Т. 32. № Web Server issue. С. W208--12.

362. Ostrow S.L. и др. Cancer Evolution Is Associated with Pervasive Positive Selection on Globally Expressed Genes // PLoS Genet. 2014. Т. 10. № 3. С. e1004239.

363. Pabo C.O., Sauer R.T. Protein-DNA recognition. // Annu Rev Biochem. 1984. Т. 53. С. 293–321.

364. Pachkov M. и др. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. // Nucleic Acids Res. 2013. Т. 41. № Database issue. С. D214-20.

365. Panne D. The enhanceosome // Curr Opin Struct Biol. 2008. Т. 18. № 2. С. 236–242.

366. Papatsenko D. ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors. // Bioinformatics. 2007. Т. 23. № 8. С. 1032–1034.

367. Papatsenko D. и др. Single-Cell Analyses of ESCs Reveal Alternative Pluripotent Cell States and Molecular Mechanisms that Control Self-Renewal // Stem Cell Reports. 2015. Т. 5. № 2. С. 207–220.

368. Papatsenko D.A., Makeev V.J. Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers // Genome …. 2002. Т. 295. № 5556. С. 813–818.

369. Pape U.J., Rahmann S., Vingron M. Natural similarity measures between position frequency matrices with an application to clustering // Bioinformatics. 2008. Т. 24. № 3. С. 350–357.

370. Papp P.P., Chattoraj D.K., Schneider T.D. Information analysis of sequences that bind the replication initiator RepA. // J Mol Biol. 1993. Т. 233. № 2. С. 219–230.

371. Paralkar V.R. и др. Unlinking an lncRNA from Its Associated cis Element. // Mol Cell. 2016. Т. 62. № 1. С. 104–10.

372. Park D. и др. Widespread misinterpretable ChIP-seq bias in yeast // PLoS One. 2013. Т. 8. № 12.

373. Park P.J. ChIP-seq: advantages and challenges of a maturing technology. // Nat Rev Genet. 2009. Т. 10. № 10. С. 669–680.

374. Parry T.J. и др. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. // Genes Dev. 2010. Т. 24. № 18. С. 2013–2018.

375. Perini G. и др. In vivo transcriptional regulation of N-Myc target genes is controlled by E-box methylation. // Proc Natl Acad Sci U S A. 2005. Т. 102. № 34. С. 12117–22.

376. Perry R.P. The architecture of mammalian ribosomal protein promoters. // BMC Evol Biol. 2005. Т. 5. С. 15.

377. Polak P. и др. Cell-of-origin chromatin organization shapes the mutational landscape of cancer // Nature. 2015. Т. 518. № 7539. С. 360–364.

378. Ponjavic J., Ponting C.P., Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs // Genome Res. 2007. Т. 17. № 5. С. 556–565.

379. Ponomarenko J. V и др. rSNP_Guide, a database system for analysis of transcription factor binding to target sequences: application to SNPs and site-directed mutations. // Nucleic Acids Res. 2001. Т. 29. № 1. С. 312–316.

380. Ponomarenko J. V и др. rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation. // Nucleic Acids Res. 2003. Т. 31. № 1. С. 118–121.

381. Portales-Casamar E. и др. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. // Nucleic Acids Res. 2010. Т. 38. № Database issue. С. D105--10.

382. Qi Y. и др. High-resolution computational models of genome binding events. // Nat Biotechnol. 2006. Т. 24. № 8. С. 963–70.

383. Quante T., Bird A. Do short, frequent DNA sequence motifs mould the epigenome? // Nat Rev Mol Cell Biol. 2016. Т. 17. № 4. С. 257–62.

384. Rach E.A. и др. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. // PLoS Genet. 2011. Т. 7. № 1. С. e1001274.

385. Ravasi T. и др. An atlas of combinatorial transcriptional regulation in mouse and man. // Cell. 2010. Т. 140. № 5. С. 744–752.

386. Ray D. и др. A compendium of RNA-binding motifs for decoding gene regulation // Nature. 2013. Т. 499. № 7457. С. 172–177.

387. Régnier M. и др. Analysis of pattern overlaps and exact computation of P-values of pattern occurrences numbers: case of Hidden Markov Models. // Algorithms Mol Biol. 2014. Т. 9. № 1. С. 25.

388. Rhee H.S., Pugh B.F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. // Cell. 2011. Т. 147. № 6. С. 1408–1419.

389. Rhee H.S., Pugh B.F. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. // Curr Protoc Mol Biol. 2012. Т. Chapter 21. С. Unit 21.24.

390. Ridinger-Saison M. и др. Spi-1/PU.1 activates transcription through clustered DNA occupancy in erythroleukemia. // Nucleic Acids Res. 2012. Т. 40. № 18. С. 8927–8941.

391. Riethoven J.-J.M. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. // Methods Mol Biol. 2010. Т. 674. С. 33–42.

392. Rishi V. и др. CpG methylation of half-CRE sequences creates C/EBPalpha binding sites that activate some tissue-specific genes. // Proc Natl Acad Sci U S A. 2010. Т. 107. № 47. С. 20311–6.

393. Rister J., Desplan C. Deciphering the genome’s regulatory code: The many languages of DNA // BioEssays. 2010. Т. 32. № 5. С. 381–384.

394. Riva A. Large-scale computational identification of regulatory SNPs with rSNP-MAPPER. // BMC Genomics. 2012. Т. 13 Suppl 4. № Suppl 4. С. S7.

395. Robertson G. и др. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. // Nat Methods. 2007. Т. 4. № 8. С. 651–657.

396. Rockel S., Geertz M., Maerkl S.J. MITOMI: A microfluidic platform for in vitro characterization of transcription factor-DNA interaction // Methods Mol. Biol. 2012. Т. 786. С. 97–114.

397. Roepcke S. и др. T-Reg comparator: An analysis tool for the comparison of position weight matrices // Nucleic Acids Res. 2005. Т. 33. № SUPPL. 2.

398. Rohs R. и др. The role of DNA shape in protein-DNA recognition. // Nature. 2009. Т. 461. № 7268. С. 1248–1253.

399. Rohs R. и др. Origins of specificity in protein-DNA recognition. // Annu Rev Biochem. 2010. Т. 79. С. 233–69.

400. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in … // Psychol Rev. 1958. Т. 65. № 6. С. 386–408.

401. Rye M.B., Sætrom P., Drabløs F. A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. // Nucleic Acids Res. 2011. Т. 39. № 4. С. e25.

402. Sanchez-Garcia F. и др. Integration of Genomic Data Enables Selective Discovery of Breast Cancer Drivers // Cell. 2014. Т. 159. № 6. С. 1461–75.

403. Sandelin A. и др. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. // Nucleic Acids Res. 2004. Т. 32. № Database issue. С. D91--4.

404. Sandelin A. и др. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. // Nat Rev Genet. 2007. Т. 8. № 6. С. 424–36.

405. SantaLucia J., Hicks D. The thermodynamics of DNA structural motifs. // Annu Rev Biophys Biomol Struct. 2004. Т. 33. С. 415–440.

406. Santra T. A bayesian framework that integrates heterogeneous data for inferring gene regulatory networks. // Front. Bioeng. Biotechnol. 2014. Т. 2. № May. С. 13.

407. Schaefer U., Schmeier S., Bajic V.B. TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. // Nucleic Acids Res. 2011. Т. 39. № Database issue. С. D106--10.

408. Schindler A.J., Sherwood D.R. The transcription factor HLH-2/E/Daughterless regulates anchor cell invasion across basement membrane in C. elegans. // Dev Biol. 2011. Т. 357. № 2. С. 380–391.

409. Schmidt D. и др. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. // Science. 2010. Т. 328. № 5981. С. 1036–40.

410. Schneider T.D. и др. Information content of binding sites on nucleotide sequences. // J Mol Biol.

1986. Т. 188. № 3. С. 415–431.

411. Schneider T.D. Consensus sequence Zen. // Appl Bioinformatics. 2002. Т. 1. № 3. С. 111–119.

412. Schneider T.D., Stephens R.M. Sequence logos: a new way to display consensus sequences. // Nucleic Acids Res. 1990. Т. 18. № 20. С. 6097–6100.

413. Schones D.E., Sumazin P., Zhang M.Q. Similarity of position frequency matrices for transcription factor binding sites. // Bioinformatics. 2005. Т. 21. № 3. С. 307–313.

414. Schubeler D. Epigenetic Islands in a Genetic Ocean // Science (80-). 2012. Т. 338. № 6108. С. 756–757.

415. Schurch N.J. и др. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? // RNA. 2016. Т. 22. № 6. С. 839–51.

416. Schwartz A.M. и др. Early B-cell factor 1 (EBF1) is critical for transcriptional control of SLAMF1 gene in human B cells // Biochim. Biophys. Acta - Gene Regul. Mech. 2016.

417. Schwartz A.M. и др. Multiple single nucleotide polymorphisms in the first intron of the IL2RA gene affect transcription factor binding and enhancer activity // Gene. 2017. Т. 602. С. 50–56.

418. Schweikert C. и др. Combining multiple ChIP-seq peak detection systems using combinatorial fusion. // BMC Genomics. 2012. Т. 13 Suppl 8. № Suppl 8. С. S12.

419. Sevignani C. и др. Mammalian microRNAs: A small world for fine-tuning gene expression // Mamm. Genome. 2006. Т. 17. № 3. С. 189–202.

420. Shelest V., Albrecht D., Shelest E. DistanceScan: a tool for promoter modeling. // Bioinformatics. 2010. Т. 26. № 11. С. 1460–1462.

421. Shen L. и др. Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters // PLoS Genet. 2007. Т. 3. № 10. С. 2023–2026.

422. Shlyueva D., Stampfel G., Stark A. Transcriptional enhancers: from properties to genome-wide predictions. // Nat Rev Genet. 2014. Т. 15. № 4. С. 272–86.

423. Siebert M., Söding J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences // Nucleic Acids Res. 2016. С. gkw521.

424. Siggers T., Gordân R. Protein-DNA binding: complexities and multi-protein codes. // Nucleic Acids Res. 2014. Т. 42. № 4. С. 2099–111.

425. Simcha D., Price N.D., Geman D. The Limits of De Novo DNA Motif Discovery // PLoS One. 2012. Т. 7. № 11. С. e47836.

426. Sinha S., Tompa M. Discovery of novel transcription factor binding sites by statistical overrepresentation. // Nucleic Acids Res. 2002. Т. 30. № 24. С. 5549–5560.

427. Slattery M. и др. Cofactor binding evokes latent differences in DNA binding specificity between hox proteins // Cell. 2011. Т. 147. № 6. С. 1270–1282.

428. Smale S.T. и др. The initiator element: A paradigm for core promoter heterogeneity within metazoan protein-coding genes // Cold Spring Harbor Symposia on Quantitative Biology. , 1998. С. 21–31.

429. Smale S.T., Kadonaga J.T. The RNA polymerase II core promoter. // Annu Rev Biochem. 2003. Т. 72. № 1. С. 449–79.

430. Smeenk L. и др. Characterization of genome-wide p53-binding sites upon stress response. // Nucleic Acids Res. 2008. Т. 36. № 11. С. 3639–54.

431. Smith Z.D., Meissner A. DNA methylation: roles in mammalian development. // Nat Rev Genet. 2013. Т. 14. № 3. С. 204–20.

432. Smits S.A., Ouverney C.C. jsPhyloSVG: A javascript library for visualizing interactive and vector-based phylogenetic trees on the web // PLoS One. 2010. Т. 5. № 8.

433. Song L., Crawford G.E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. // Cold Spring Harb Protoc. 2010. Т. 2010. № 2. С. pdb.prot5384.

434. Soufi A. и др. Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming // Cell. 2015. Т. 161. № 3. С. 555–568.

435. Spirin S. и др. NPIDB: a database of nucleic acids-protein interactions. // Bioinformatics. 2007. Т. 23. № 23. С. 3247–8.

436. Stampfel G. и др. Transcriptional regulators form diverse groups with context-dependent regulatory functions. // Nature. 2015. Т. 528. № 7580. С. 147–51.

437. Steensel B. van, Delrow J., Henikoff S. Chromatin profiling using targeted DNA adenine methyltransferase. // Nat Genet. 2001. Т. 27. № 3. С. 304–308.

438. Stegmaier P. и др. A Discriminative Approach for Unsupervised Clustering of DNA Sequence Motifs // PLoS Comput Biol. 2013. Т. 9. № 3.

439. Stegmaier P., Kel A.E., Wingender E. Systematic DNA-binding domain classification of transcription factors. // Genome Inform. 2004. Т. 15. № 2. С. 276–286.

440. Stormo G.D. и др. algorithm to distinguish translational initiation sites in E. coli // Nucleic Acids Res. 1982. Т. 10. С. 2997–3011.

441. Stormo G.D. DNA binding sites: representation and discovery. // Bioinformatics. 2000. Т. 16. № 1. С. 16–23.

442. Stormo G.D. Introduction to Protein-DNA Interactions: Structure, Thermodynamics, and Bioinformatics. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press, 2013. Вып. 1. 218 с.

443. Stormo G.D., Schneider T.D., Gold L. Quantitative analysis of the relationship between nucleotide sequence and functional activity // Nucleic Acids Res. 1986. Т. 14. С. 6661–6679.

444. Straussman R. и др. Developmental programming of CpG island methylation profiles in the human genome. // Nat Struct Mol Biol. 2009. Т. 16. № 5. С. 564–571.

445. Struhl K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. // Nat Struct Mol Biol. 2007. Т. 14. № 2. С. 103–105.

446. Strunnikova M. и др. Chromatin inactivation precedes de novo DNA methylation during the progressive epigenetic silencing of the RASSF1A promoter. // Mol Cell Biol. 2005. Т. 25. № 10. С. 3923–3933.

447. Suryamohan K., Halfon M.S. Identifying transcriptional cis-regulatory modules in animal genomes. // Wiley Interdiscip. Rev. Dev. Biol. 2015. Т. 4. № 2. С. 59–84.

448. Szalkowski A.M., Schmid C.D. Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. // Brief Bioinform. 2011. Т. 12. № 6. С. 626–33.

449. Taatjes D.J., Marr M.T., Tjian R. Regulatory diversity among metazoan co-activator complexes. // Nat Rev Mol Cell Biol. 2004. Т. 5. № May. С. 403–410.

450. Takahashi K., Yamanaka S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors // Cell. 2006. Т. 126. № 4. С. 663–676.

451. Talarek N., Bontron S., Virgilio C. De. Quantification of mRNA stability of stress-responsive yeast genes following conditional excision of open reading frames. // RNA Biol. 2013. Т. 10. № 8. С. 1299–308.

452. Tanaka E. и др. Improved similarity scores for comparing motifs // Bioinformatics. 2011. Т. 27. № 12. С. 1603–1609.

453. Tantin D. и др. High-throughput biochemical analysis of in vivo location data reveals novel distinct classes of POU5F1(Oct4)/DNA complexes // Genome Res. 2008. Т. 18. № 4. С. 631–639.

454. Tate P.H., Bird A.P. Effects of DNA methylation on DNA-binding proteins and gene expression // Curr Opin Genet Dev. 1993. Т. 3. № 2. С. 226–231.

455. Teng M. и др. Regsnps: A strategy for prioritizing regulatory single nucleotide substitutions // Bioinformatics. 2012. Т. 28. № 14. С. 1879–1886.

456. Terada N. Rapamycin Selectively Inhibits Translation of mRNAs Encoding Elongation Factors and Ribosomal Proteins // Proc. Natl. Acad. Sci. 1994. Т. 91. № 24. С. 11477–11481.

457. Teytelman L. и др. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. // Proc Natl Acad Sci U S A. 2013. Т. 110. № 46. С. 18602–7.

458. The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource

(UniProt). // Nucleic Acids Res. 2012. Т. 40. № Database issue. С. D71--5.

459. Thomas-Chollier M. и др. RSAT: regulatory sequence analysis tools. // Nucleic Acids Res. 2008. Т. 36. № Web Server issue.

460. Thomas-Chollier M. и др. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. // Nat Protoc. 2012. Т. 7. № 8. С. 1551–1568.

461. Thomas M.C., Chiang C.-M. The general transcription machinery and general cofactors. // Crit Rev Biochem Mol Biol. 2006. Т. 41. № 3. С. 105–178.

462. Thomas R. и др. Features that define the best ChIP-seq peak calling algorithms // Brief Bioinform. 2016. С. bbw035.

463. Thompson J.D., Gibson T.J., Higgins D.G. Multiple sequence alignment using ClustalW and ClustalX // Curr Protoc Bioinforma. 2002. Т. Chapter 2. С. Unit 2 3.

464. Thoreen C.C. и др. A unifying model for mTORC1-mediated regulation of mRNA translation. // Nature. 2012. Т. 485. № 7396. С. 109–113.

465. Thurman R.E. и др. The accessible chromatin landscape of the human genome. // Nature. 2012. Т. 489. № 7414. С. 75–82.

466. Tomazou E.M., Meissner A. Epigenetic regulation of pluripotency // Adv Exp Med Biol. 2010. Т. 695. С. 26–40.

467. Tomovic A., Oakeley E.J. Position dependencies in transcription factor binding sites // Bioinformatics. 2007. Т. 23. № 8. С. 933–941.

468. Tompa M. и др. Assessing computational tools for the discovery of transcription factor binding sites. // Nat Biotechnol. 2005. Т. 23. № 1. С. 137–144.

469. Topisirovic I., Sonenberg N. mRNA translation and energy metabolism in cancer: the role of the MAPK and mTORC1 pathways. // Cold Spring Harb Symp Quant Biol. 2011. Т. 76. С. 355–367.

470. Touzet H., Varré J.-S. Efficient and accurate P-value computation for Position Weight Matrices. // Algorithms Mol Biol. 2007. Т. 2. С. 15.

471. Tran N.T.L., Huang C.-H. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. // Biol Direct. 2014. Т. 9. С. 4.

472. Trcek T. и др. Single-molecule mRNA decay measurements reveal promoter-regulated mRNA stability in yeast // Cell. 2011. Т. 147. № 7. С. 1484–1497.

473. Trinklein N.D. и др. Identification and functional analysis of human transcriptional promoters // Genome Res. 2003. Т. 13. № 2. С. 308–312.

474. Tronche F. и др. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome // J Mol Biol. 1997. Т. 266. № 2. С. 231–245.

475. Tuerk C., Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. // Science. 1990. Т. 249. № 4968. С. 505–510.

476. Tuteja G. и др. Extracting transcription factor targets from ChIP-Seq data. // Nucleic Acids Res. 2009. Т. 37. № 17. С. e113.

477. Valouev A. и др. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. // Nat Methods. 2008. Т. 5. № 9. С. 829–834.

478. Vaquerizas J.M. и др. A census of human transcription factors: function, expression and evolution. // Nat Rev Genet. 2009. Т. 10. № 4. С. 252–263.

479. Varshney A. и др. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. // Proc Natl Acad Sci U S A. 2017. С. 201621192.

480. Verfaillie A. и др. iRegulon and i-cisTarget: Reconstructing Regulatory Networks Using Motif and Track Enrichment. // Curr Protoc Bioinformatics. 2015. Т. 52. С. 2.16.1-2.16.39.

481. Verfaillie A. и др. Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic // Genome Res. 2016. Т. 26. № 7. С. 882–895.

482. Vernot B. и др. Personal and population genomics of human regulatory variation. // Genome Res. 2012. Т. 22. № 9. С. 1689–97.

483. Viré E. и др. The Polycomb group protein EZH2 directly controls DNA methylation. // Nature. 2006. Т. 439. № 7078. С. 871–874.

484. Vorontsov I.E. и др. PERFECTOS-APE: Predicting regulatory functional effect of SNPs by approximate P-value estimation // Proceedings of the BIOINFORMATICS 2015 - 6th International Conference on Bioinformatics Models, Methods and Algorithms. Lisbon, Portugal: SCITEPRESS, 2015. С. 102–108.

485. Vorontsov I.E., Kulakovskiy I. V, Makeev V.J. Jaccard index based similarity measure to compare transcription factor binding site models. // Algorithms Mol Biol. 2013. Т. 8. № 1. С. 23.

486. Vorontsov I.E.I.E. и др. Negative selection maintains transcription factor binding motifs in human cancer // BMC Genomics. 2016. Т. 17. С. 395.

487. Wagner A. A computational genomics approach to the identification of gene networks. // Nucleic Acids Res. 1997. Т. 25. № 18. С. 3594–3604.

488. Wallerman O. и др. Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. // Nucleic Acids Res. 2009. Т. 37. № 22. С. 7498–508.

489. Walsh C.P., Bestor T.H. Cytosine methylation and mammalian development. // Genes Dev. 1999. Т. 13. № 1. С. 26–34.

490. Wang G., Yu T., Zhang W. WordSpy: identifying transcription factor binding motifs by building a dictionary and learning a grammar. // Nucleic Acids Res. 2005. Т. 33. № Web Server issue. С. W412--6.

491. Wang H. и др. Widespread plasticity in CTCF occupancy linked to DNA methylation // Genome Res. 2012a. Т. 22. № 9. С. 1680–1688.

492. Wang J. и др. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors // Genome Res. 2012b. Т. 22. № 9. С. 1798–1812.

493. Wang J. и др. Factorbook.org: A Wiki-based database for transcription factor-binding data generated by the ENCODE consortium // Nucleic Acids Res. 2013a. Т. 41. № D1.

494. Wang R. и др. LOcating non-unique matched tags (LONUT) to improve the detection of the enriched regions for ChIP-seq data. // PLoS One. 2013b. Т. 8. № 6. С. e67788.

495. Wang X., Proud C.G. mTORC1 signaling: what we still don’t know. // J Mol Cell Biol. 2011. Т. 3. № 4. С. 206–220.

496. Warren C.L. и др. Defining the sequence-recognition profile of DNA-binding molecules. // Proc Natl Acad Sci U S A. 2006. Т. 103. № 4. С. 867–872.

497. Wasserman W.W., Sandelin A. Applied bioinformatics for the identification of regulatory elements. // Nat Rev Genet. 2004. Т. 5. № 4. С. 276–287.

498. Weber M. и др. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. // Nat Genet. 2007. Т. 39. № 4. С. 457–66.

499. Wederell E.D. и др. Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing // Nucleic Acids Res. 2008. Т. 36. № 14. С. 4549–4564.

500. Wei C.-L. и др. A global map of p53 transcription-factor binding sites in the human genome. // Cell. 2006. Т. 124. № 1. С. 207–219.

501. Weirauch M.T. и др. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity // Cell. 2014. Т. 158. № 6. С. 1431–1443.

502. Whitaker J.W., Chen Z., Wang W. Predicting the human epigenome from DNA motifs. // Nat Methods. 2015. Т. 12. № 3. С. 265–272.

503. Whitfield T.W. и др. Functional analysis of transcription factor binding sites in human promoters. // Genome Biol. 2012. Т. 13. № 9. С. R50.

504. Whitington T. и др. Inferring transcription factor complexes from ChIP-seq data. // Nucleic Acids Res. 2011. Т. 39. № 15. С. e98.

505. Wilbanks E.G. и др. A workflow for genome-wide mapping of archaeal transcription factors with ChIP-seq. // Nucleic Acids Res. 2012. Т. 40. № 10. С. e74.

506. Wilbanks E.G., Facciotti M.T. Evaluation of algorithm performance in ChIP-seq peak detection. //

PLoS One. 2010. Т. 5. № 7. С. e11471.

507. Wingender E. и др. TRANSFAC: an integrated system for gene expression regulation. // Nucleic Acids Res. 2000. Т. 28. № 1. С. 316–319.

508. Wingender E. Criteria for an updated classification of human transcription factor DNA-binding domains. // J Bioinform Comput Biol. 2013. Т. 11. № 1. С. 1340007.

509. Wingender E. и др. TFClass: a classification of human transcription factors and their rodent orthologs. // Nucleic Acids Res. 2015. Т. 43. № Database issue. С. D97-102.

510. Wingender E., Schoeps T., Donitz J. TFClass: an expandable hierarchical classification of human transcription factors // Nucleic Acids Res. 2012. Т. 41. № D1. С. D165–D170.

511. Wolberger C. Combinatorial transcription factors // Curr Opin Genet Dev. 1998. Т. 8. № 5. С. 552–559.

512. Wong D. и др. Extensive characterization of NF-κB binding uncovers non-canonical motifs and advances the interpretation of genetic functional traits. // Genome Biol. 2011. Т. 12. № 7. С. R70.

513. Wu F., Olson B.G., Yao J. DamID-seq: Genome-wide Mapping of Protein-DNA Interactions by High Throughput Sequencing of Adenine-methylated DNA Fragments // J. Vis. Exp. 2016. № 107. С. 1–11.

514. Wu J. и др. The initiator motif is preferentially used as the core promoter element in ionizing radiation-responsive genes. // Radiat Res. 2006. Т. 166. № 5. С. 810–3.

515. Wyers F. и др. Cryptic Pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase // Cell. 2005. Т. 121. № 5. С. 725–737.

516. Xie X., Rigor P., Baldi P. MotifMap: a human genome-wide map of candidate regulatory motif sites. // Bioinformatics. 2009. Т. 25. № 2. С. 167–174.

517. Xing H. и др. Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. // PLoS Comput Biol. 2012. Т. 8. № 7. С. e1002613.

518. Xu B. и др. A structural-based strategy for recognition of transcription factor binding sites. // PLoS One. 2013. Т. 8. № 1. С. e52460.

519. Xu M., Su Z. A novel alignment-free method for comparing transcription factor binding site motifs // PLoS One. 2010. Т. 5. № 1.

520. Yamashita R. и др. Comprehensive detection of human terminal oligo-pyrimidine (TOP) genes and analysis of their characteristics. // Nucleic Acids Res. 2008. Т. 36. № 11. С. 3707–3715.

521. Yamashita R. и др. DBTSS: DataBase of Transcriptional Start Sites progress report in 2012. // Nucleic Acids Res. 2012. Т. 40. № Database issue. С. D150--4.

522. Yan J. и др. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. // Cell. 2013. Т. 154. № 4. С. 801–813.

523. Yang C. и др. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. // Gene. 2007. Т. 389. № 1. С. 52–65.

524. Yang J. и др. A novel SALL4/OCT4 transcriptional feedback network for pluripotency of embryonic stem cells // PLoS One. 2010. Т. 5. № 5.

525. Yang L. и др. TFBSshape: A motif database for DNA shape features of transcription factor binding sites // Nucleic Acids Res. 2014. Т. 42. № D1.

526. Yilmaz A. и др. AGRIS: The arabidopsis gene regulatory information server, an update // Nucleic Acids Res. 2011. Т. 39. № SUPPL. 1.

527. Yokoyama K.D., Ohler U., Wray G.A. Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships. // Nucleic Acids Res. 2009. Т. 37. № 13. С. e92.

528. Young R.A. Control of the embryonic stem cell state. // Cell. 2011. Т. 144. № 6. С. 940–54.

529. Yusuf D. и др. The transcription factor encyclopedia. // Genome Biol. 2012. Т. 13. № 3. С. R24.

530. Zahnow C.A. CCAAT/enhancer-binding protein beta: its role in breast cancer and associations with receptor tyrosine kinases. // Expert Rev Mol Med. 2009. Т. 11. С. e12.

531. Zandevakili P., Hu M., Qin Z. GPUmotif: An ultra-fast and energy-efficient motif analysis program using graphics processing units // PLoS One. 2012. Т. 7. № 5.

532. Zanegina O. и др. An updated version of NPIDB includes new classifications of DNA-protein complexes and their families. // Nucleic Acids Res. 2016. Т. 44. № D1. С. D144-53.

533. Zantoni M. и др. Motif Discovery Fixing Mismatch Positions. // Commun. to SIMAI Congr. 2007. Т. 1.

534. Zaret K.S., Carroll J.S. Pioneer transcription factors: establishing competence for gene expression. // Genes Dev. 2011. Т. 25. № 21. С. 2227–41.

535. Zhang Y. и др. Model-based analysis of ChIP-Seq (MACS). // Genome Biol. 2008. Т. 9. № 9. С. R137.

536. Zhao Y., Granas D., Stormo G.D. Inferring binding energies from selected binding sites // PLoS Comput Biol. 2009. Т. 5. № 12.

537. Zhao Y., Stormo G.D. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. // Nat Biotechnol. 2011. Т. 29. № 6. С. 480–3.

538. Zhu L.J. и др. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. // BMC Bioinformatics. 2010. Т. 11. С. 237.

539. Zhu L.J. и др. FlyFactorSurvey: A database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system // Nucleic Acids Res. 2011. Т. 39. № SUPPL. 1.

540. Zid B.M., O'Shea E.K. Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast. TL - 514 // Nature. 2014. Т. 514 VN-. № 7520. С. 117–121.

541. Zoncu R., Efeyan A., Sabatini D.M. mTOR: from growth signal integration to cancer, diabetes and ageing. // Nat Rev Mol Cell Biol. 2011. Т. 12. № 1. С. 21–35.