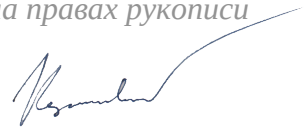
на правах рукописи



Кулаковский Иван Владимирович

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

03.01.09 – «Математическая биология, биоинформатика»

Автореферат диссертации на соискание ученой степени
доктора биологических наук

Москва – 2017

Работа выполнена в Лаборатории вычислительных методов системной биологии Федерального государственного бюджетного учреждения науки Института молекулярной биологии им. В.А. Энгельгардта Российской академии наук (ИМБ РАН), Федеральное агентство научных организаций.

Официальные оппоненты:

Орлов Юрий Львович, доктор биологических наук, профессор РАН,
Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук» (ИЦиГ СО РАН)

Озолинь Ольга Николаевна, доктор биологических наук, профессор,
Федеральное государственное бюджетное учреждение науки Институт биофизики клетки Российской академии наук (ИБК РАН)

Галзитская Оксана Валерьевна, доктор физико-математических наук,
Федеральное государственное бюджетное учреждение науки Институт белка Российской академии наук (ИБ РАН)

Ведущая организация:

Институт математических проблем биологии РАН – филиал Федерального государственного учреждения «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук» (ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН).

Защита диссертации состоится «___» _____ 201_ г. на заседании диссертационного совета по защите докторских и кандидатских диссертаций Д002.077.04 (утвержден Приказом Минобрнауки России от 16 декабря 2013 года №978/нк) при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН) по адресу: 127051, г. Москва, Большой Каретный переулок, д.19 стр. 1., факс: +7 (495) 650-05-79, e-mail: gir@iitp.ru

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН) и на сайте института: http://iitp.ru/ru/dissertation/1340.htm

Автореферат разослан «___» _____ 201_ г.

Ученый секретарь совета
д.б.н., профессор Г.И. Рожкова

Общая характеристика работы

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.

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.

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.

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[1], 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.

Апробация и публикации по теме работы

Список публикаций по теме диссертации включает **21** статью в рецензируемых международных журналах, **2** приглашенные главы-обзора, **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), «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).

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).

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,

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

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.

Структура и объем работы

Диссертация построена по стандартной схеме с тремя ключевыми разделами: «Обзор литературы», «Материалы и методы», «Результаты и обсуждение». Список литературы содержит 541 наименование. Объем диссертации составляет 245 страниц, включая 48 рисунков и 6 таблиц.

Материалы и методы

В этом разделе рассматривается вычислительное представление мотивов и авторские методы идентификации и сравнения мотивов.

Позиционно-весовая матрица как модель мотива

У высших эукариот сайты связывания факторов транскрипции сравнительно коротки (10-20 п.н.), а их мотивы «вырождены», т.е. допускают неточные совпадения. Задача идентификации мотива, т.е. паттерна, описывающего область частичного локального сходства между последовательностями, часто формулируется как построение множественного безделеционного локального выравнивания [Stormo, 2000]. Высоко консервативные позиции выравнивания (с предпочтительным содержанием конкретных нуклеотидов) соответствуют позициям сайтов

связывания, которые являются более важными для ДНК-белкового узнавания. Стандартным способом представления мотива являются позиционно-весовая матрица (ПВМ): для мотива сайтов связывания длины m нуклеотидов ПВМ представляет собой матрицу $4\times m$, где строки соответствуют нуклеотидам {**A**, **C**, **G**, **T**}, а столбцы – позициям сайта связывания. Числа в матрице – веса – соответствуют предпочтению или избеганию конкретного нуклеотида в конкретной позиции. Каждому олигонуклеотиду-слову длины m на ДНК-алфавите ПВМ сопоставляет оценку – вещественное число, которое получается как произведение или сумма весов в соответствующих ячейках матрицы. В классической работе [Berg, von Hippel, 1988] было показано, что при правильном подборе весов ПВМ-оценка слов пропорциональна аффинности сайтов. Мы традиционно строим ПВМ по выравниванию как в работе [Lifanov и др., 2003]. Для визуализации выравниваний и мотивов сегодня повсеместно используются лого-диаграммы [Schneider, Stephens, 1990]: по горизонтальной оси откладываются позиции, а по вертикальной оси информационное содержание колонки выравнивания (на основе энтропии Шэннона), причем высоты отдельных нуклеотидов соответствуют их относительным частотам.

Статистическая значимость вхождений мотивов

Шкала ПВМ-оценок зависит от формальных параметров: длины мотива и способа преобразования частот в веса. В большинстве случаев удобнее явная единая шкала оценок. Одним из способов получения такой шкалы является перевод оценок в статистические значимости. Каждому слову ПВМ сопоставляет вещественное 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 фракция словаря, состоящая из слов с оценками не хуже t и под P-значением мотива для порога t мы понимаем объем этой доли словаря. Иначе говоря, P-значение соответствует площади под правым хвостом распределения оценок ПВМ, т.е. вероятности случайно выбрать из словаря слово с оценкой не хуже t. Для получения точного P-значения в работе [Touzet, Varré, 2007] было предложено использовать динамическое программирование. Мы реализовали этот алгоритм [Vorontsov и др., 2015] для вычисления P-значений как обычных, так и динуклеотидных весовых матриц (см. ниже).

Сравнение мотивов с помощью операционных характеристик приемника

Выбор порогового значения оценки превращает весовую матрицу в бинарный классификатор «да/нет»: принадлежит ли конкретное слово на ДНК-алфавите множеству потенциальных сайтов связывания. Эта классификация не является статической и зависит от порога оценки.

Достаточно часто для сравнения мотивов в смысле точности распознавания сайтов используются стандартные инструменты машинного обучения и анализа сигналов, в частности, кривая операционной характеристики приемника (receiver operating characteristic, ROC). ROC-кривая строится в координатах «доля истинных положительных предсказаний» (ось Y, чувствительность) и «доля ложных положительных предсказаний» (ось X, доля ошибок первого рода, 1 – специфичность). В качестве свободного параметра выступает пороговое значение оценки, т.е. каждая точка на ROC-кривой соответствует конкретному порогу деления ДНК-слов на «сайты» и «не сайты». Площадь под ROC-кривой (area under ROC-curve, AUC ROC) удобно использовать в качестве количественного значения качества классификации во всем диапазоне пороговых значений.

Корректный подсчет числа истинных положительных предсказаний возможен на основе

позитивного контроля – независимой выборки экспериментально определенных сайтов связывания. Это достижимо либо с помощью кросс-валидации данных различных экспериментов, либо путем деления одной выборки на поднаборы для обучения и тестирования. Для подсчета ложноположительных предсказаний необходима информация об участках ДНК, достоверно не связываемых белком. Прямые экспериментальные данные такого типа чаще всего недоступны, и в качестве негативного контроля обычно используют близлежащие геномные районы [Agius и др., 2010] или случайные последовательности [Keilwagen, Grau, 2015]. В нашей работе для этой цели используется вероятность случайного нахождения сайта связывания для всевозможных последовательностей заданной длины [Kulakovskiy и др., 2013]. Для каждого порогового значения оценки мотива мы подсчитываем P_s как вероятность случайной встречи хотя бы 1 надпорогового вхождения ПВМ в случайную последовательность ДНК фиксированной длины L, выбранной как медиана длин последовательностей позитивного контроля. P_s подсчитывается на основе P-значения мотива как вероятность получения ПВМ-оценки не хуже порога для случайного слова хотя бы в одном месте случайной двуцепочечной последовательности ДНК, в предположении, что вхождения ПВМ (включая перекрывающиеся вхождения) являются независимыми, и их полное число удовлетворяет составному распределению Пуассона: $P_s = 1 - (1 - P)^{2(L-m+1)}$, где m –длина мотива, а P-значение мотива (P), может подсчитываться с учетом нуклеотидного или динуклеотидного состава. Если нуклеотиды равновероятны, то P_s – доля всевозможных последовательностей длины L, в которых находится хотя бы одно надпороговое вхождение мотива.

Идентификация мотивов в больших выборках нуклеотидных последовательностей с учетом априорной информации

Современные экспериментальные методы требуют новых биоинформатиеских подходов и не только вследствие растущего объема данных. Так, замена гибридизации на микрочипах секвенированием в методах на основе иммунопреципитации хроматиина (переход от ChIP-chip к ChIP-Seq) снабдила современные полногеномные данные по взаимодействиям белок-ДНК принципиально важной дополнительной информацией – «формой пиков», т.е. позиционным профилем покрытия сегментов прочтениями. Первые методы [Hu и др., 2010], явно использующие форму пиков ChIP-Seq, имели недостаточную производительность, использовали сложные форматы входных данных и давали нестабильные результаты, что ограничивало их практическое применение. Мы предложили новую модификацию алгоритма ChIPMunk [Kulakovskiy и др., 2010], являющуюся развитием ранее опубликованного нами алгоритма [Kulakovskiy, Makeev, 2009] и лишенную этих недостатков, что было подтверждено и в независимых тестах [Bi и др., 2011; Kuttippurathu и др., 2011; Ma и др., 2012].

Алгоритм ChIPMunk является методом «жадной» оптимизации и последовательно решает две задачи: (а) построение псевдо-оптимального безделеционного локального выравнивания последовательностей, определяющего 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 тестирования самосогласованности весовой матрицы. Получаемые ChIPMunk выравнивания не являются действительно оптимальными в силу стохастического алгоритма, основанного на случайном сэмплировании обучающей выборки. В то же время, поиск действительно оптимального выравнивания не обязателен, поскольку выборка имеющихся сайтов связывания всегда является каким-то подмножеством

«истинных возможных» сайтов.

Среди множества возможных выравниваний ChIPMunk выбирает мотив с наилучшим дискретным информационным содержанием [Hertz, Stormo, 1999]. Мы используем дискретное информационное содержание с кульбаковским членом (КДИС):

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

где m – длина мотива, q_α – частота наиболее редкого нуклеотида α (для простоты при масштабировании КДИС фиксирована как 0.25), N – полное число слов в выравнивании, $i \epsilon \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$ – 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.

Для быстрого поиска кандидатных мотивов ChIPMunk использует случайное сэмплирование учебной выборки. Возврат к полной выборке для весовых матриц, частично оптимизированных на подвыборке, позволяет получить полное выравнивание всех последовательностей.

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. Аналогичным образом ChIPMunk учитывает и буквы в нотации IUPAC. Идея взвешивания отсчетов $x_{\alpha,j}$ используется и для учета позиционных профилей последовательностей, например, информации о покрытии прочтениями ChIP-Seq:

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

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

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

Здесь 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, индикатор δ является символом Кронекера, а $\mathbf{N}$ – неизвестный «любой» нуклеотид, отсчеты которого затем равномерно распределяются между четырьмя реальными. Таким образом, позиционные профили явным образом штрафуют информационное содержание мотивов, локализованных в своих низких участках, за счет «размывания» значений отсчетов конкретных нуклеотидов. Для подсчета оценок при выборе наилучшего вхождения в последовательностях с профилями используется следующая формула:

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

Распределение информационного содержания по колонкам выравнивания сайтов связывания факторов транскрипции часто имеет периодичность, схожую с размерами витка спирали ДНК [Papp, Chattoraj, Schneider, 1993]. 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 [Schneider, 2002].

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 получения слабовыраженного паттерна. Форма мотива является вектором длины m (длина мотива) 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). Форма мотива влияет только на выбор наилучшего слова в каждой последовательности, но не меняет перестройку весовой матрицы по выравниванию лучших слов. ChIPMunk может использовать два варианта «формы» мотива: однобоксовую и двухбоксовую ($\cos^2(\pi n/T)$ и $\sin^2(\pi n/T)$, где $T = 10.5$ и соответствует шагу спирали ДНК, а n является относительной координатой в списке колонок, т.е. $n = 0$ в центре мотива).

ChIPMunk может обрабатывать различные типы экспериментальных данных и подходит для анализа сайтов связывания факторов транскрипции не только в эукариотических, но и в прокариотических геномах. Анализ нескольких сотен последовательностей может быть проведен на персональном компьютере за несколько минут. ChIPMunk был интегрирован в несколько программных комплексов, предназначенных для анализа мотивов и результатов высокопроизводительного секвенирования, в частности, BioUML[2], MotifLab[3] и Nebula[4].

Построение расширенных моделей мотивов с учетом корреляций соседних позиций

Предположение о независимости соседних позиций сайтов связывания является одним из основных ограничений при использовании классических позиционно-весовых матриц. Прямым расширением ПВМ является алогичная весовая матрица большей размерности, учитывающая корреляции соседних нуклеотидов [Benos, Bulyk, Stormo, 2002]. Алгоритм diChIPMunk предназначени для построения динуклеотидных позиционно-весовых матриц (диПВМ) и полностью повторяет ChIPMunk, но все операции совершаются на динуклеотидном алфавите. Каждый динуклеотид состоит из двух соседних нуклеотидов, и обратно, каждый нуклеотид, кроме первого и последнего в последовательности, принадлежит двум перекрывающимся динуклеотидам. 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. Для определения оптимальности выравнивания diChIPMunk, как и ChIPMunk, пользуется дискретным информационным содержанием, переформулированным для динуклеотидного алфавита.

Для тестирования ChIPMunk и diChIPMunk мы сравнивали ROC-кривые для известных мотивов и мотивов, найденных *de novo* в данных ChIP-Seq проекта ENCODE. Пример сравнения мотива из TRANSFAC и результатов ChIPMunk (ПВМ) и diChIPMunk (диПВМ) приведен на **Рисунке 1**. В качестве альтернативного теста мы изучили локализацию сайтов связывания в

---

[2] BioUML Platform. http://biouml.org

[3] MotifLab is a general workbench for analysing regulatory sequence regions. http://motiflab.org

[4] Nebula web service. http://nebula.curie.fr/

данных ChIP-Seq, предполагая, что более достоверные сайты должны быть идентифицированы ближе к вершинам пиков, т.е. в позициях с наилучшим покрытием прочтениями.

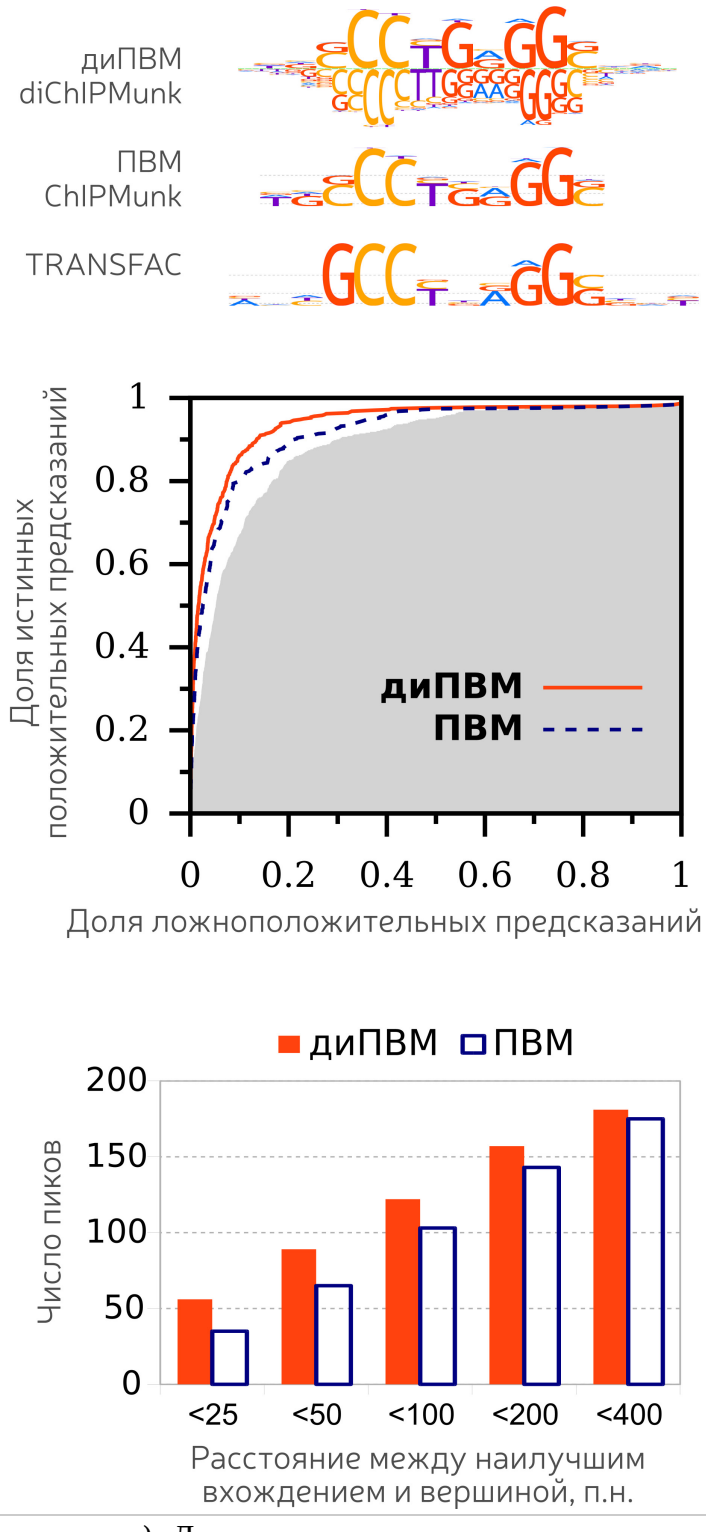


Рисунок 1. *(верхняя панель)* Лого-диаграммы мотивов связывания фактора AP2. Для динуклеотидной ПВМ дополнительно показаны частоты динуклеотидов, захватывающих пары соседних колонок выравнивания. (*средняя панель*) ROC-кривые мотивов, полученных ChIPMunk (ПВМ, пунктирная линия) и diChIPMunk (диПВМ, сплошная линия), и мотива из TRANSFAC (выделена серая площадь под ROC-кривой). (*нижняя панель*) Кумулятивное распределение расстояний между наилучшими вхождениями мотивов и вершинами пиков. Рассмотрены только пики с несовпадающими (на расстоянии не менее чем 5 п.н.) наилучшими вхождениями ПВМ и диПВМ. Рисунок адаптирован из работы [Kulakovskiy и др., 2013].

В заметном числе последовательностей наилучшие вхождения ПВМ и диПВМ оказывались на удалении друг от друга более чем на 5 п.н, для таких случаев мы рассмотрели локализацию наилучших предсказанных сайтов в различных окнах вокруг наивысшей точки позиционного профиля – вершины пика (<25 п.н. до вершины, <50 п.н. до вершины, и т.д). И действительно, наилучшие с точки зрения диПВМ сайты оказались сфазированы точнее.

Дополнительно, на примере фактора FoxA2 было проведено сравнительное тестирование различных моделей мотивов на независимых ChIP-Seq данных с подбором порогов оценок на основании результатов гель-шифт экспериментов [Levitsky и др., 2014]. В этом тестировании наилучшим оказался мотив, построенный diChIPMunk (**Рисунок 2**).

Наконец, для широкого спектра факторов транскрипции результаты diChIPMunk были систематически верифицированы путем кросс-валидации на независимых ChIP-Seq данных в ходе построения базы данных HOCOMOCO (см. «Результаты и обсуждение»).

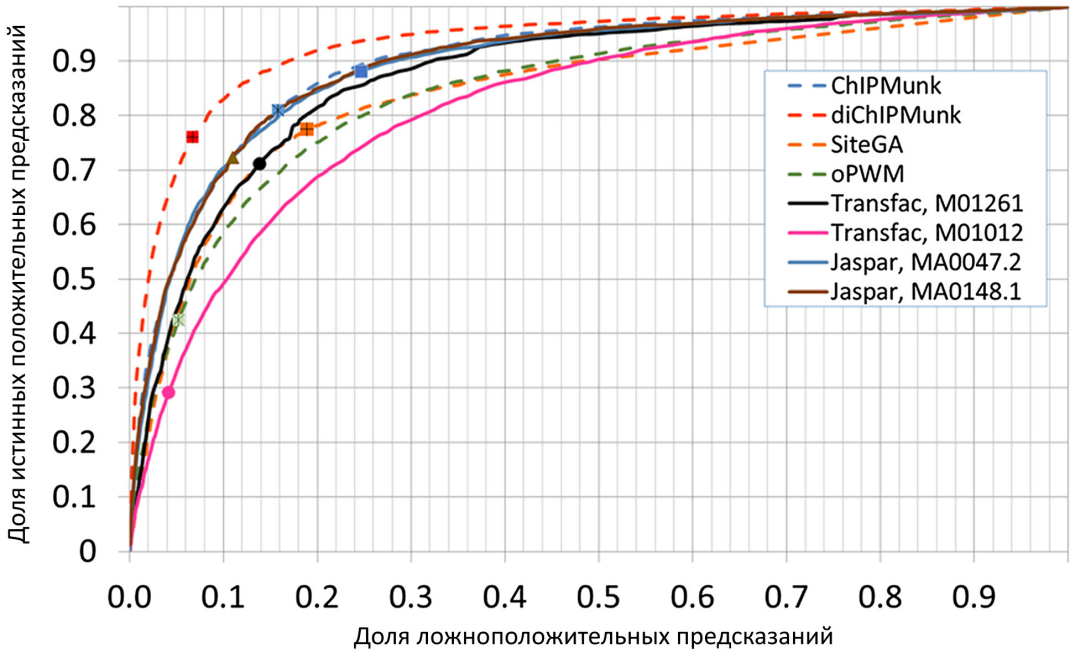


Рисунок 2. ROC-кривые для различных моделей мотивов связывания фактора FoxA2. Для построения мотивов ChIPMunk и diChIPMunk использованы данные по связыванию в печени взрослой мыши [Wederell и др., 2008]. Данные ChIP-Seq в клетках HepG2 [Wallerman и др., 2009] использованы в качестве положительного контроля выборки. Модели SiteGA и oPWM построены в ИЦиГ на основе догеномных экспериментальных данных. В сравнение включены матрицы из TRANSFAC и JASPAR. Маркеры на графиках соответствуют порогам оценок, определенным по сайтам, верифицированным с помощью гель-шифт эксперимента. Рисунок адаптирован из работы [Levitsky и др., 2014].

Естественная мера сходства мотивов

Позиционно-весовая матрица с пороговым значением оценки задает модель паттерна: ранжирование слов и определение «наилучшего» подмножества. Мотивы, идентифицированные различными вычислительными методами в различных экспериментальных данных для одного и того же белка, часто похожи, но почти никогда не являются идентичными; это верно и для изучения мотивов связывания белков одного структурного семейства. Таким образом, возникает задача сравнения мотивов с точки зрения прямого сходства паттернов. Стандартные методы

сравнения ПВМ 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 работе [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 при учете частот динуклеотидов. Алгоритм и программа носят название MACRO-APE (MAtrix CompaRisOn by Approximate P-value Estimation, сравнение матриц с использованием приблизительной оценки *P*-значений).

Поиск вхождений мотивов

Для большинства практических задач необходим поиск вхождений мотивов в заданных последовательностях, для мононуклеотидных матриц существует множество готовых инструментов. Для расширенных моделей, учитывающих зависимости между нуклеотидами, быстрые методы поиска вхождений появились только недавно [Korhonen и др., 2016]. Мы разработали свой инструмент, подходящий для поиска вхождений моно- и динуклеотидных весовых матриц: SPRy-SARUS (Straightforward yet Powerful Rapid SuperAlphabet Representation Utilized for motif Search, супералфавитное представление для поиска мотивов). Этот простой метод использует супералфавитный подход [Korhonen и др., 2009], который позволяет уменьшить число операций при последовательном сканировании последовательности путем замены алфавита: переход от моно- к динуклеотидам, или от ди- к тринуклеотидам. Это отличается от подхода, использованного в diChIPMunk для диПВМ: соседние нуклеотиды в супералфавитной записи не перекрываются, а выигрыш по времени достигается за счет манипуляции с представлением матрицы, которая в случае супералфавита содержит оценки не для отдельных букв, а сразу суммарные для пар букв, что вдвое уменьшает эффективную длину мотива и необходимое число операций при подсчете оценок слов.

Аннотация однонуклеотидных вариантов

Генетические варианты, например однонуклеотидные полиморфизмы, могут быть ассоциированы с экспрессией генов. Одним из механизмов, обеспечивающих эту связь, может быть колокализация вариантов и сайтов связывания факторов транскрипции. Объяснить или предсказать изменение аффинности сайтов в зависимости от конкретных аллелей можно с помощью анализа мотивов. В частности, *P*-значение мотива для ПВМ-оценки конкретного сайта зависит от аллельных вариантов и их локализации относительно наиболее консервативных позиций мотива, а отношение *P*-значений может использоваться как аналог изменения энергии

связывания, вызванного однонуклеотидной заменой [Berg, 1987]. Мы реализовали подсчет и сравнение *P*-значений для однонуклеотидных вариантов в программе PERFECTOS-APE (Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation), которая по постановке повторяет существующие работы [Manke, Heinig, Vingron, 2010], но в качестве моделей мотивов позволяет использовать и моно-, и динуклеотидные позиционно-весовые матрицы.

Результаты и обсуждение

Раздел посвящен описанию основных результатов диссертации, полученных путем применения авторских вычислительных методов. 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.

Коллекция HOCOMOCO: мотивы сайтов связывания факторов транскрипции человека и мыши

Существует множество репозиториев-коллекций известных мотивов сайтов ДНК-связывающих белков, которые можно использовать при анализе регуляторных последовательностей. Проблема практического использования коллекций состоит в «несистематической избыточности». Для конкретного фактора транскрипции даже в одной коллекции часто присутствует несколько похожих мотивов, построенных по различным экспериментальным данным и/или различными вычислительными методами. В этом разделе диссертации обсуждается построение коллекции HOCOMOCO (изначально – HOmo sapiens COmprehensive MOtif Collection), содержащей мотивы связывания для сотен факторов транскрипции человека и мыши. Мотивы в HOCOMOCO были построены единообразным способом с помощью авторских программ 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 (например, в виде мономера и димера) подтверждены экспериментально.

Внешнее ограничение степеней свободы позволяет повысить устойчивость процедуры выявления мотива. При построении HOCOMOCO мы использовали возможности программы ChIPMunk: (1) прямую интеграцию данных различных экспериментальных методов [Kulakovskiy, Makeev, 2009]; (2) априорную информацию о вероятном положении сайта связывания (например, о форме пиков ChIP-Seq); (3) априорную информацию о форме самого мотива (вписывая одно- и двухбоксовые мотивы в один и два витка спирали ДНК). Результаты вычислительной идентификации мотивов затем проходили стадию экспертного курирования.

В первой версии коллекции мы сделали акцент на интеграции данных о сайтах связывания из различных источников, включая систематизированные базы данных JASPAR и TRANSFAC, из которых были собраны данные о регионах связывания, аннотированные по литературным данным. Для первого публичного релиза HOCOMOCO также использовались предварительные результаты ChIP-Seq проекта ENCODE и первые массовые результаты HT-SELEX [Jolma и др., 2010]. В ходе экспертного курирования для каждого фактора транскрипции выделяли наиболее релевантную модель среди всех, построенных ChIPMunk, и одновременно

производилась экспертная оценка достоверности мотива от A (наилучший) до D (наименее достоверный) по нескольким критериям, учитывая распределение информационного содержания по колонкам и сходство мотивов с известными и между членами одного семейства.

Первый публичный релиз v9 коллекции HOCOMOCO (HOmo sapiens COmprehensive MOtif COllection) содержал 426 мотивов 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., модели качества A и B были построены по наборам последовательностей, в среднем объединяющим два и более источника данных. Используя ChIP-Seq данные ENCODE позволил провести тестирование качества для мотивов связывания лишь нескольких десятков факторов транскрипции. Мы сравнили мотивы TRANSFAC, JASPAR и HOCOMOCO путем оценки площади под ROC-кривыми и выяснили, что мотивы из HOCOMOCO оказались более точными в 90% случаев.

Расширение коллекции путем систематического анализа данных ChIP-Seq

Основной вклад в коллекцию HOCOMOCO v9 внесли «догеномные» данные, но наиболее удачные модели были построены на основе данных ChIP-Seq. Возникла естественная мотивация расширить коллекцию с фокусом на данных высокопроизводительного секвенирования и провести полномасштабное сравнительное тестирование позиционно-весовых матриц. 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 [Yevshin и др., 2017]. Использование GTRD, в частности, позволило выделить отдельную группу мотивов связывания для факторов транскрипции мыши.

Также, для нескольких сотен факторов транскрипции были опубликованы данные HT-SELEX [Jolma и др., 2013] по связыванию *in vitro*. Пользуясь достаточным объемом данных, мы приняли решение дополнительно построить расширенные динуклеотидные позиционно-весовые матрицы на основе данных HT-SELEX и ChIP-Seq и включить такие модели в общее сравнительное тестирование.

Общая идея построения коллекции сохранилась: мы использовали ChIPMunk для идентификации мотивов и курировали результаты на предмет согласованности мотивов с уже известными в рамках одного семейства. В то же время, данные ChIP-Seq позволили провести систематическое тестирование качества распознавания сайтов связывания в данных ChIP-Seq, не использованных при идентификации мотивов.

Использованный релиз GTRD содержал результаты 1690 экспериментов и покрывал 392 (96) фактора транскрипции для человека (мыши), учитывая только эксперименты с 200 и более идентифицированными регионами связывания в геноме. Для каждого эксперимента регионы были ранжированы по высоте пика (т.е. по максимальному покрытию чтениями), не более 1000 наилучших были взяты для анализа. Регионы с четными рангами использовались для идентификации мотивов с помощью ChIPMunk и diChIPMunk, регионы с нечетными рангами использовались как положительный контроль при тестировании.

Все мотивы, идентифицированные в ChIP-Seq данных, были проаннотированы вручную: лишь половина прошла этап курирования и затем участвовала в автоматизированном сравнительном тестировании (692 мотивов для человека плюс 177 для мыши из полного множества 1690 наборов последовательностей). Для проведения сравнительного тестирования

мы использовали несколько публичных коллекций мотивов: JASPAR, HOMER, SwissRegulon, исходные мотивы, определенные авторами HT-SELEX, и предыдущую сборку HOCOMOCO. Чтобы сравнить точность новых и существующих моделей мы использовали AUC ROC. В силу априорно неизвестной достоверности исходных данных и точности известных мотивов необходимо было одновременно выявить как неудачные эксперименты ChIP-Seq, так и неверные мотивы. Для каждого фактора транскрипции это делалось путем подсчета взвешенного среднего AUC ROC по всем наборам пиков, где в качестве веса выступает средний AUC конкретного набора пиков для всех мотивов. Чтобы избежать систематических искажений от ошибочных мотивов и наборов пиков, мы итеративно убирали из общей выборки наборы пиков и мотивы с взвешенным средним AUC < 0.65. Пороговое значение 0.65 было выбрано из соображения, что модели HOCOMOCO v9 наиболее высокого качества A и B не должны были покинуть сравнительное тестирование. Такая процедура, в частности, позволила для каждого фактора транскрипции с ChIP-Seq данными выбрать наилучший мотив с наибольшим взвешенным средним AUC и использовать это значение для автоматического назначения рейтинга качества (аналогично HOCOMOCO v9 – от A до D, учитывая помимо AUC и число проанализированных экспериментальных источников). Дополнительно мы сохраняли альтернативные мотивы под качеством S. Коллекция HOCOMOCO v10 была собрана из новых мотивов, полученных на основе ChIP-Seq, и мотивов HOCOMOCO v9. По сравнению с предыдущим релизом, HOCOMOCO v10 дополнительно покрывает мотивами еще две сотни факторов транскрипции человека, и содержит более сотни динуклеотидных моделей.

Суммарно, HOCOMOCO v10 включает 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.

Коллекция HOCOMOCO доступна и в машинно-читаемом виде, и в человеко-читаемом виде через веб-интерфейс по адресу http://hocomoco.autosome.ru. Основным идентификатором фактора транскрипции в HOCOMOCO является ключ по базе UniProt[5], но предоставляются ссылки и на другие ключевые биоинформатические базы данных (в том числе Entrez Gene[6], HGNC[7], MGI[8], FANTOM5 SSTAR[9]). В HOCOMOCO v10 мы явным образом включили информацию о структурных семейства ДНК-связывающих доменов согласно классификации TFClass[10]. На **Рисунке 3** представлено иерархическое дерево, показывающее подсемейства и семейства факторов транскрипции и их покрытие мотивами в HOCOMOCO v10.

---

[5] UniProt. http://www.uniprot.org/

[6] Home - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene

[7] HGNC database of human genes. http://www.genenames.org

[8] MGI - Mouse Genome Informatics. http://www.informatics.jax.org/

[9] FANTOM5_SSTAR. http://fantom.gsc.riken.jp/5/sstar/Main_Page

[10] TFClass: Classification of Human Transcription Factors and Mouse Orthologs. http://tfclass.bioinf.med.uni-goettingen.de

Для сравнительной оценки общего качества мотивов в различных коллекциях для каждой коллекции мы подсчитали число факторов транскрипции, мотивы которых показали наилучший средний взвешенный AUC среди всех коллекций. Тестирование подтвердило высокое качество мотивов HOCOMOCO (**Рисунок 4**). На сегодня публикации о HOCOMOCO собрали уже десятки цитирований, в том числе, коллекция упомянута как ключевой информационный ресурс в фундаментальном обзоре по изучению геномных вариантов в сайтах связывания [Deplancke и др., 2016] и использована командами-победителями международного соревнования ENCODE-DREAM по вычислительному предсказанию сайтов связывания факторов транскрипции.

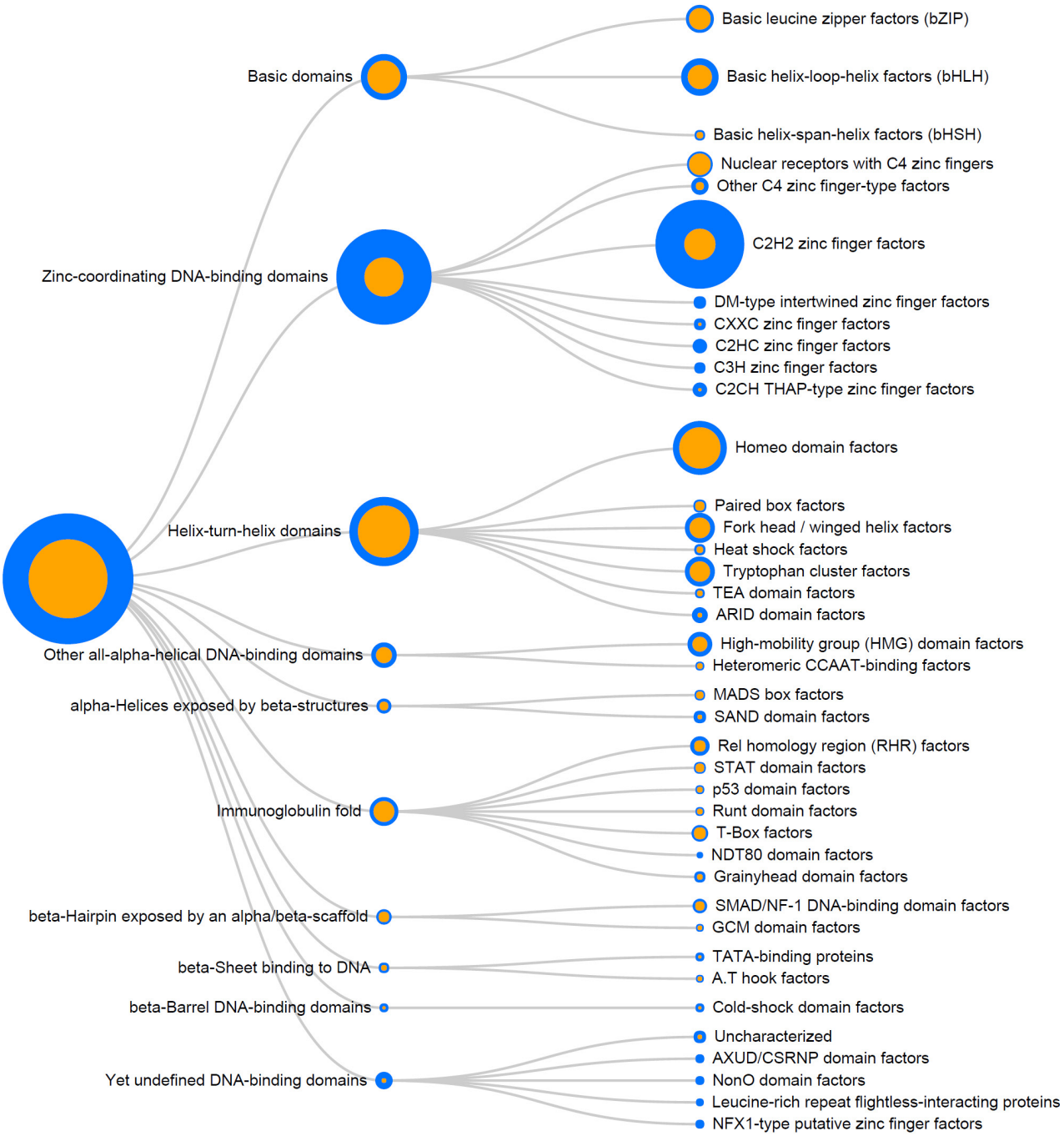


Рисунок 3. Покрытие основных структурных семейств факторов транскрипции моделями HOCOMOCO v10. Площадь синих кругов пропорциональна полному числу членов конкретного семейства; оранжевые внутренние круги показывают долю факторов транскрипции, для которых доступны мотивы связывания. Классификация факторов транскрипции дана по TFClass [Wingender и др., 2015]. Рисунок адаптирован из работы [Kulakovskiy и др., 2016].

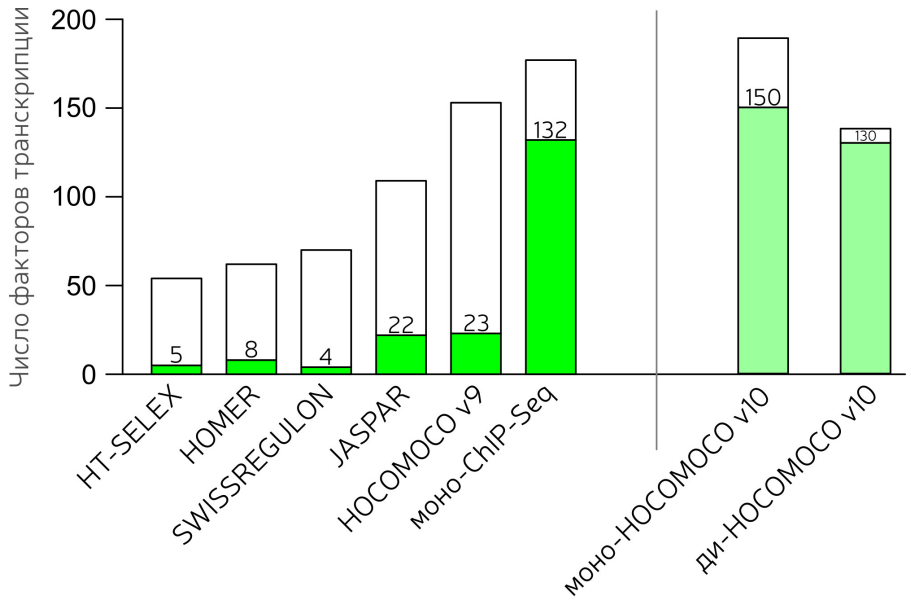


Рисунок 4. Результаты тестирования качества мотивов в различных коллекциях на основе данных ChIP-Seq для факторов транскрипции мыши и человека. Полная высота полос показывает полное число факторов транскрипции, мотивы для которых были протестированы в конкретном случае. Зеленая часть полос соответствует числу факторов транскрипции, мотивы которых оказались наилучшими. Белая часть полосы для мононуклеотидных позиционно-весовых матриц HOCOMOCO v10 соответствует факторам транскрипции, для которых наилучшие модели не были построены ChIPMunk и не вошли в итоговую коллекцию (т.е. принадлежали коллекциям HOMER, SWISSREGULON, JASPAR и опубликованным ранее моделям HT-SELEX). Рисунок адаптирован из работы [Kulakovskiy и др., 2016].

Практический анализ мотивов в избранных регуляторных системах

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

Возможность контролируемой настройки механизмов контроля самообновления клеток и плюрипотентности является ключевым вопросом регенеративной биологии. 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 [Takahashi, Yamanaka, 2006]. В работе [Papatsenko и др., 2015] был проведен высокопроизводительный анализ экспрессии нескольких десятков генов в сотнях отдельных эмбриональных стволовых клеток мыши. Анализ паттернов экспрессии позволил установить наличие двух характерных субпопуляций клеток и реконструировать участок генной сети, потенциально обеспечивающей устойчивые альтернативные состояния. Дмитрием Папаценко было предложено существование некогерентной петли прямой-обратной связи [Goentoro и др., 2009], включающей OCT4 и NANOG и предполагающей антагонизм этих двух ключевых регуляторов. Интересно, как антагонизм OCT4 и NANOG может реализовываться с точки зрения структуры регуляторных последовательностей. Для прояснения этого вопроса мы провели детальный анализ имеющихся в открытом доступе ChIP-Seq данных и известных мотивов связывания для OCT4, SOX2 и NANOG.

Полногеномный профиль связывания факторов плюрипотентности в эмбриональных стволовых клетках мыши и человека исследовался с помощью метода ChIP-Seq в нескольких

работах. Результаты части экспериментов были позднее единообразно обработаны Дж. Гоке (J. Göke), который любезно поделился результатами. Для каждого эксперимента мы извлекли 1000 наиболее достоверных 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. Идентификация мотивов показала стабильные результаты как при использовании различных источников данных, так и при сравнении «человек-мышь». Мотивы ChIPMunk и diChIPMunk на основе данных [Chen и др., 2008] в среднем оказались наилучшими и были взяты в сравнение с ранее известными. Результирующие ROC-кривые для мотивов NANOG представлены на **Рисунке 5**.

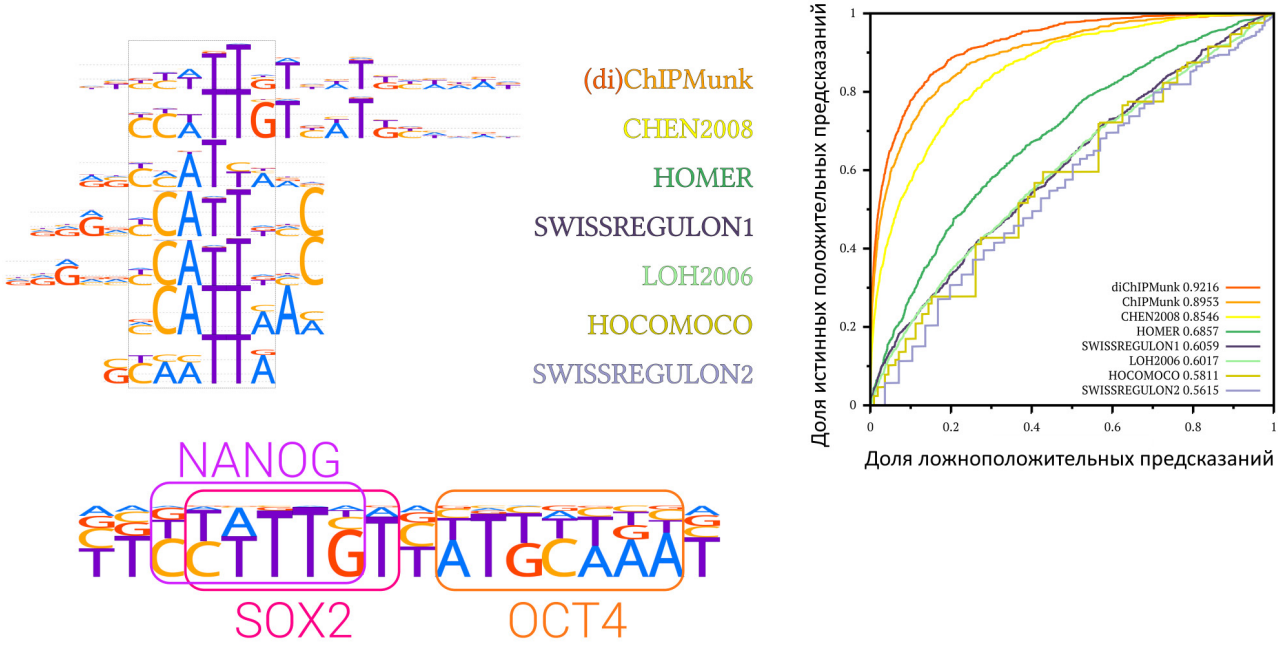


Рисунок 5. Сравнение моделей мотивов фактора транскрипции NANOG и предложенная схема тройственного композитного элемента. (*левая панель*) Мотивы связывания NANOG и (*правая панель*) ROC-кривые для сравнения качества распознавания сайтов в данных ChIP-Seq. Значения AUC ROC приведены в легенде. (*нижняя панель*) Предлагаемый консенсус тройственного композитного элемента OCT4-SOX2/NANOG. Рисунок адаптирован из работы [Papatsenko и др., 2015].

Для OCT4 и SOX2 *de novo* мотивы представляли 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 *de novo* мотив для NANOG во всех случаях повторял парный мотив OCT4-SOX2. Более того, среди всех мотивов NANOG только похожим на OCT4-SOX2 удается успешно распознать сайты связывания *in vivo*. По всей видимости, в регуляторных районах, связываемых NANOG, присутствует только небольшое число его «истинных» собственных сайтов, а большинство регионов связывания содержат композитный элемент, соответствующий перекрытию сайтов нескольких регуляторов. Мы сформулировали идею тройственного OSN-композитного элемента OCT4-SOX2/NANOG, в котором NANOG узнает участок, перекрывающийся с боксом SOX2, и, вероятно, препятствует

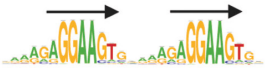
устойчивому связыванию гетеродимера OCT4-SOX2. Это согласуется с антагонизмом OCT4 и NANOG в модели генной сети, и подтверждается независимыми исследованиями совместного участия OCT4 и NANOG в регуляции экспрессии [Bin Le и др., 2014]. Интересно, что факт перекрытия сайтов SOX2 и NANOG в композитном элементе с OCT4 был описан ранее для дистального энхансера OCT4 [Young, 2011]. Кроме того, NANOG и SOX2 способны самостоятельно взаимодействовать без участия OCT4 и связывать собственный характерный композитный элемент, похожий, но не идентичный OSN-мотиву [Gagliardi и др., 2013]. Таким образом, в реальности мы имеем дело с интерференцией большего числа сигналов, и декомпозиция мотивов в ChIP-Seq для ключевых факторов плюрипотентности продолжает представлять интерес.

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

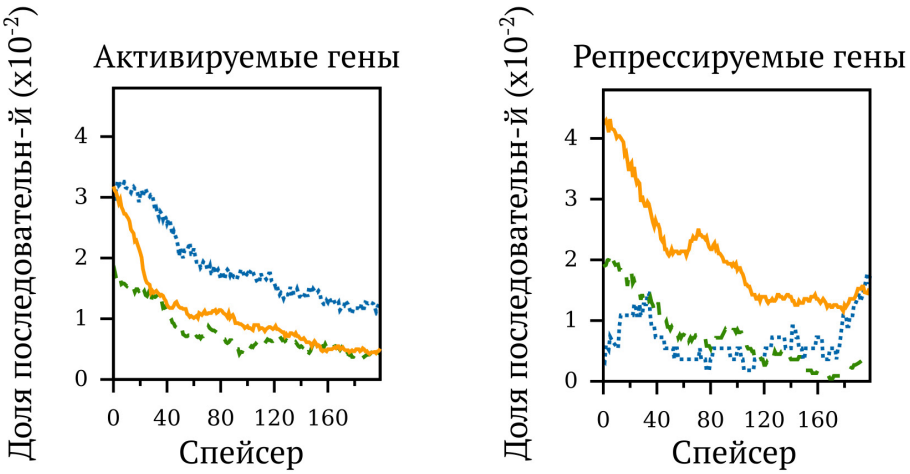
Фактор транскрипции Spi1/PU.1 является одним из ключевых регуляторов экспрессии генов в клетках костного мозга и B-лимцофитах [Iwasaki и др., 2005]. Оверэкспрессия Spi1 блокирует дифференцировку предшественников эритроцитов и приводит к эритролейкемии [Moreau-Gachelin и др., 1996]. Spi1 относится к ETS-семейству факторов транскрипции и связывает характерный консенсус с коровым элементом GGAA. С помощью комбинации ChIP-Seq и экспрессионного профилирования на микрочипах [Ridinger-Saison и др., 2012] нашим коллегам из института Кюри (Париж) удалось не только установить полногеномный профиль сайтов связывания, но и соотнести связывание фактора транскрипции с изменением экспрессии близлежащих генов. Наш вклад в 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 (более 17 тысяч регионов) и хорошо согласуется с литературными данными. Затем пики были разбиты на группы: (а) в соответствии с геномной локализацией с фокусом на проксимальных промоторах, включая ближайшие 3’-окрестности стартов транскрипции (от -1.5 тыс. п.н. в 5’ области до +2 тыс. п.н.), и потенциальных энхансерах (вплоть до -30 тыс. п.н. до старта); (б) в соответствии с предполагаемой функцией (активация либо ингибирование экспрессии ближайшего гена). Фиксированный порог на изменение экспрессии в полтора раза выявил 672 гена, активируемых и ингибируемых при сравнении результатов, полученных при оверэкспрессии и нокдауне Spi1, среди них более 70% локусов включали пики Spi-1 в окрестности от -30 тыс. п.н. в 5’ область до 5 тыс. п.н. в 3’ область относительно старта гена.

Задача по анализу мотивов состояла в определении структуры регуляторной последовательности, определяющей результат связывания Spi1: активация или ингибирование экспрессии генов-мишеней. Мы изучили позиционные предпочтения связывания Spi1 в пиках ChIP-Seq путем поиска вхождений мотива с порогом оценки на уровне *P*-значения 0.0001(6), что соответствует случайным предсказаниям примерно в 10% двуцепочечных последовательностей длины 300 п.н., близкой к характерной протяженности пиков ChIP-Seq. В результате удалось обнаружить замечательный факт: «знак» эффекта, оказываемого Spi1, зависит от взаимной ориентации парных сайтов связывания (**Рисунок 6**). Связывание одинаково ориентированных тандемных сайтов в промоторах чаще ведет к активации экспрессии, но только для промоторов, не пересекающихся с CpG-островками (вероятно, это связано с типом промоторов или характерными кофакторов). Обратно, связывание Spi1 тандемных сайтов в энхансерных областях приводит к ингибированию экспрессии. Нельзя полностью исключить, что парные

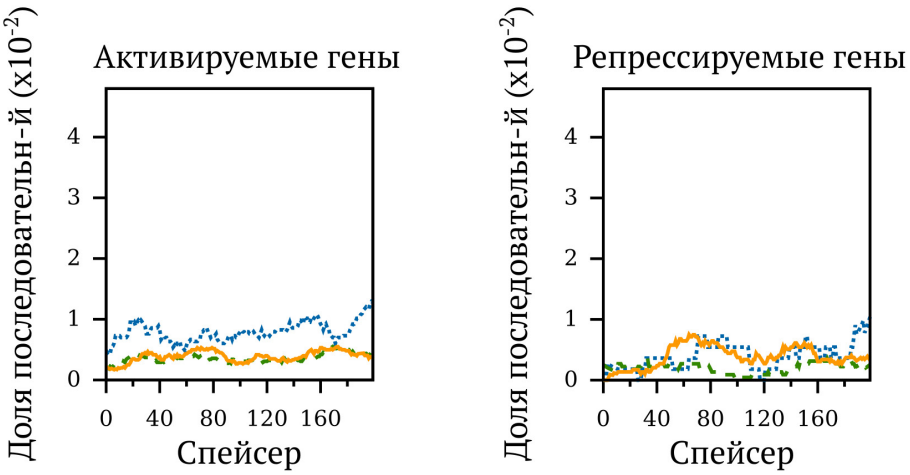
сайты в каких-то случаях соответствуют посадке и других членов ETS-семейства помимо Spi1, но нам кажется замечательным сам факт различной функциональной роли гомотипических пар сайтов с одинаковой ориентацией в зависимости от геномной локализации. Интересно, что пары сайтов с противоположной ориентацией 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



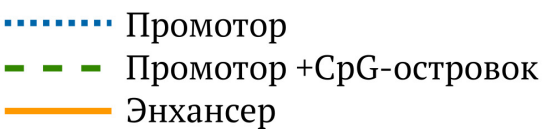


Рисунок 6. Предпочтительные расстояния между парами сайтов Spi1 в пиках ChIP-Seq в тандеме (*верхняя панель*) и в обратно-комплементарной ориентации (*нижняя панель*).

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].

Взаимосвязь транскрипции и трансляции мРНК-мишеней каскада mTOR

Киназа mTORC1 является одним из ключевых регуляторов клеточного роста и пролиферации у высших эукариот [Wang, Proud, 2011]. Каскад mTOR играет важную роль в онкогенезе [Topisirovic, Sonenberg, 2011], что стимулирует исследования молекулярных механизмов, в том числе, роли каскада mTOR в регуляции трансляции.

Особенности 5’ нетранслируемых областей (НТО) мРНК, трансляция которых зависит от клеточного роста, изучается давно. Более двадцати лет назад был обнаружен короткий 5’-Терминальный ОлигоПиримидиновый мотив (ТОП), задействованный в регуляции трансляции рибосомных белков [Jefferies, Thomas, 1994; Terada, 1994]. Сегодня известно, что эволюционно консервативный ТОП содержат многие мРНК рибосомных белков и факторов трансляции [Perry, 2005], а мотив представляет собой олигопиримидиновую последовательность длины 5-14 нуклеотидов, локализованную непосредственно на 5’ конце [Meyuhas, 2000]. В свою очередь, на уровне транскрипции был выявлен специальный класс промоторов, зависимых от TCT-мотива, строго локализованного в позиции, где непосредственно стартует инициация транскрипции [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.

Данные рибосомного профайлинга (Ribo-Seq) по изменению эффективности трансляции при ингибировании сигнального каскада mTOR повторно подняли вопрос о необходимости ТОП мотива для регуляции трансляции. На основании анализа сотен мРНК-мишеней mTOR была предположена функциональность нетерминальных и вырожденных мотивов ТОП [Thoreen и др., 2012] и внутренних олигопиримидиновых трактов в удалении от 5’ конца 5’ НТО [Hsieh и др., 2012]. В то же время, локализацию мотивов относительно 5’ конца мРНК невозможно выяснить без точной позиционной информации об инициации транскрипции. Эта информация стала доступной благодаря кэп-анализу экспрессии генов (CAGE) и технологии Helicos для секвенирования одной молекулы без использования амплификации ПЦР [Kanamori-Katayama и др., 2011]. То есть, появилась возможность объединить точные данные по трансляции (Ribo-Seq) с данными по транскрипции (CAGE), чтобы прояснить вопрос о ДНК и РНК-мотивах, участвующих в mTOR-опосредованной регуляции на уровнях транскрипции и трансляции. В работе [Eliseeva и др., 2013] мы применили методы анализа мотивов, чтобы заново идентифицировать мотив ТОП и оценить его локализацию относительно 5’ концов 5’ НТО и стартов транскрипции.

С помощью ChIPMunk в 5’ НТО мРНК-мишеней mTOR мы выявили основной CU-богатый мотив, для этого использовали 250 транскриптов 142 генов-трансляционных мишеней mTOR человека из работы [Hsieh и др., 2012]. Оптимальный мотив длины 14 хорошо согласуется с известным ранее представлением ТОП с мажорным цитозином в UCU(TCT) контексте [Yamashita и др., 2008]. По аналогии с ТОП-мотивом в мРНК мы называем соответствующий участок ДНК ОлигоПиримидиновым мотивом (ОП).

Чтобы выяснить точную локализацию олигопиримидиновых мотивов относительно стартов транскрипции, мы подсчитывали число 5’ НТО с ОП-вхождениями в конкретной позиции относительно максимума CAGE-сигнала. Далее мы оценили статистическую значимость ассоциации между принадлежностью транскрипта множеству mTOR-мишеней (используя прочие транскрипты в качестве контроля) и наличию вхождений ОП-мотива на конкретной позиции 5’ НТО. 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, выбранного для ОП-мотива. Обогащения ОП-мотива в 3’ области относительно старта обнаружено не было. Таким образом, нам не удалось найти аргументов ни в пользу существования альтернативных мотивов, похожих на ТОП [Thoreen и др., 2012], ни в пользу обогащения пиримидин-богатыми трактами «глубин» 5’ НТО [Hsieh и др., 2012].

Гены рибосомных белков и основных факторов трансляции транскрибируются с TCT-промоторов, для которых характерна точная инициация транскрипции и, как следствие, наличие ТОП-мотива непосредственно на 5’ конце мРНК. В то же время, многие mTOR-мишени транскрибируются с широких промоторов. Количественные данные CAGE позволяют оценить ширину региона инициации транскрипции («ширину старта» гена) как минимальную протяженность региона, покрывающего старты как минимум 2/3 пула конкретной изоформы мРНК. Выяснилось, что треть мишеней mTOR транскрибируется с широких стартов (шириной более 10 п.н.). Идентификация ТОП-мотива отдельно для широких стартов выявила подтип, похожий на «внутренний» пиримидин-богатый мотив из работы [Hsieh и др., 2012]. Таким образом, гипотеза о функциональной роли внутренних пиримидин-богатых участков по всей видимости является продуктом неточной аннотации широких стартов инициации транскрипции.

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, для которых только часть региона покрыта олигопиримидиновым трактом (**Рисунке 7**). Локальное переключение активности стартов на транскрипционном уровне может влиять на присутствие ТОП в мРНК [Kleene и др., 2003], то есть регуляция транскрипции может определять дальнейшую регуляцию трансляции.

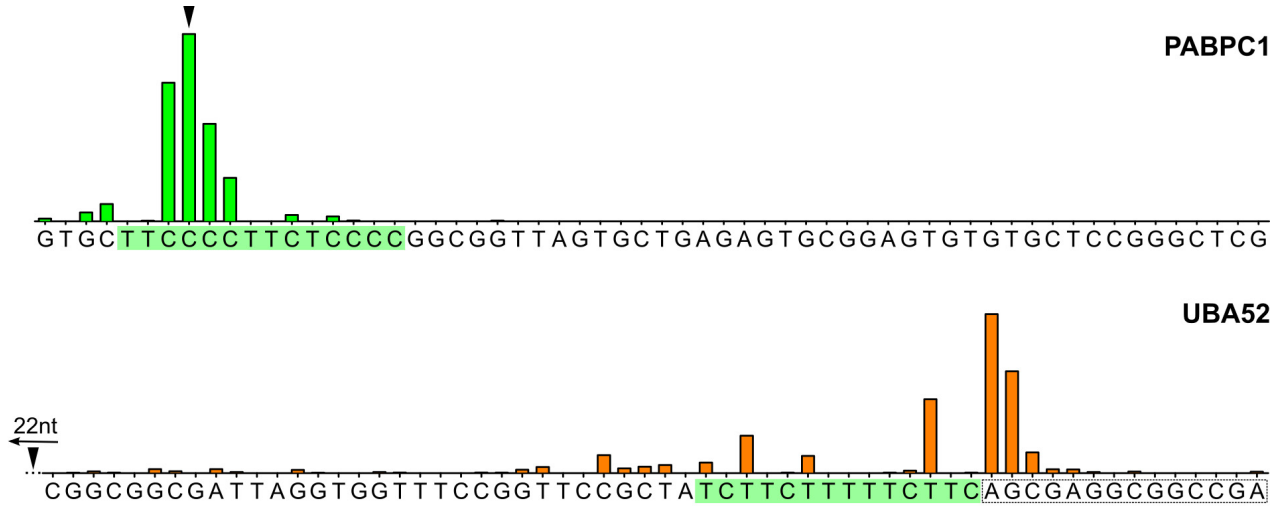


Рисунок 7. CAGE-сигнал и вхождения ОП-мотива в расширенные 5’ НТО трех мРНК-мишеней mTOR: PABPC1 (*верхняя панель*), UBA52 (*нижняя панель*). Черным треугольником показан старт транскрипции согласно аннотации генома человека (UCSC hg18). Наилучшие вхождения ОП-мотива в последовательность выделены цветным фоном. Рисунок адаптирован из работы [Eliseeva и др., 2013].

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

Соматические мутации в сайтах связывания факторов транскрипции 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. Объем информации о роли некодирующих вариантов в онкогенезе будет расти вместе с объемом данных по индивидуальной геномике. Например, в недавнем исследовании ассоциации геномных вариантов с риском развития эпителиального рака яичников [Lawrenson и др., 2015] лишь 2 полиморфизма из почти 300 значимых оказались локализованы в кодирующих областях, при этом 25 перекрывались с сайтами связывания факторов транскрипции.

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

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. Из полного набора в несколько десятков миллионов картированных мутаций мы выбрали только изменения в потенциальных регуляторных областях (интронах и промоторах), которые составляли примерно половину общей выборки. Затем мы картировали потенциальные сайты связывания факторов транскрипции (используя модели HOCOMOCO и *P*-значения 0.0005) 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 фланках 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 (учитывались изменения не менее чем в 4 раза). Для оценки давления отбора мы сравнивали наблюдаемые частоты мутаций, значительно меняющих предсказанную аффинность, с ожидаемыми частотами, оцененными по контрольным данным: (1) «перемешанный» контроль, состоящий из последовательностей со случайно переставленными нуклеотидами; (2) геномный контроль, полученный случайным сэмплированием сегментов промоторов и интронов, не перекрывающихся с окнами, центрированными на реальных мутациях. Чтобы учесть систематический вклад мутационных подписей, характерных для различных типов раковых клеток, предсказанные в контрольных данных сайты связывания сэмплировались, чтобы в итоге получить одинаковые распределения мутационных контекстов (исходный нуклеотид, мутация и пара ближайших фланкирующих нуклеотидов) для реальных данных и контроля для каждого фактора транскрипции.

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.

Изменения, вызываемые мутациями чаще, чем ожидается, находятся под положительным отбором. Среди факторов транскрипции, сайты которых в результате мутаций теряют аффинность значительно чаще, чем ожидается, присутствовали члены семейств AP2 и C/EBP, что согласуется с опубликованными данными [Khurana и др., 2013]. Мутации в сайтах связывания других белков, в частности, членов семейства ETS, чаще, чем ожидается, вызывают усиление аффинности (т.е. создание сайта). Для значительно более широкого спектра мотивов мутации, приводящие к изменению аффинности сайтов, избегаются в различных типах рака. В частности, мутаций избегают сайты связывания факторов транскрипции, принадлежащих классу ядерных рецепторов. Кроме того, под отрицательным отбором находятся сайты белков семейств HOX и FOX, что обнаружено и для нормальных клеток [Vernot и др., 2012].

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 8** показано распределение замен во вхождениях мотива связывания AP2A (часто повреждаемых мутациями) и ESR1 (избегают мутагенеза) для рака молочной железы. Нормализованные частоты замен отображены параллельно лого-визуализации мотива: 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. Одновременно, по сравнению с контролем частота мутаций G(+4) в геноме рака молочной железы выше в 1.5 раза. Контрастная картина наблюдается в мажорной позиции C(+4) в контексте TCA мотива связывания эстрогенового рецептора ESR1: мутации наблюдаются существенно реже по сравнению с любым из контролей. Не менее наглядно сравнение следующих двух боксов TGA в мотиве ESR1: 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, судя по всему, менее важен для связывания. В этой позиции аккумулируется значительно больше соматических мутаций.

Чтобы увеличить достоверность предсказаний сайтов связывания, мы дополнительно рассматривали поднаборы мутаций, соответствующие ДНКазо-доступным участкам [Thurman и др., 2012] в промоторах и интронах. 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.

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 сайтов белкового семейства. Поиск таких специфических случаев может помочь выявлению критических участков регуляторных сетей.

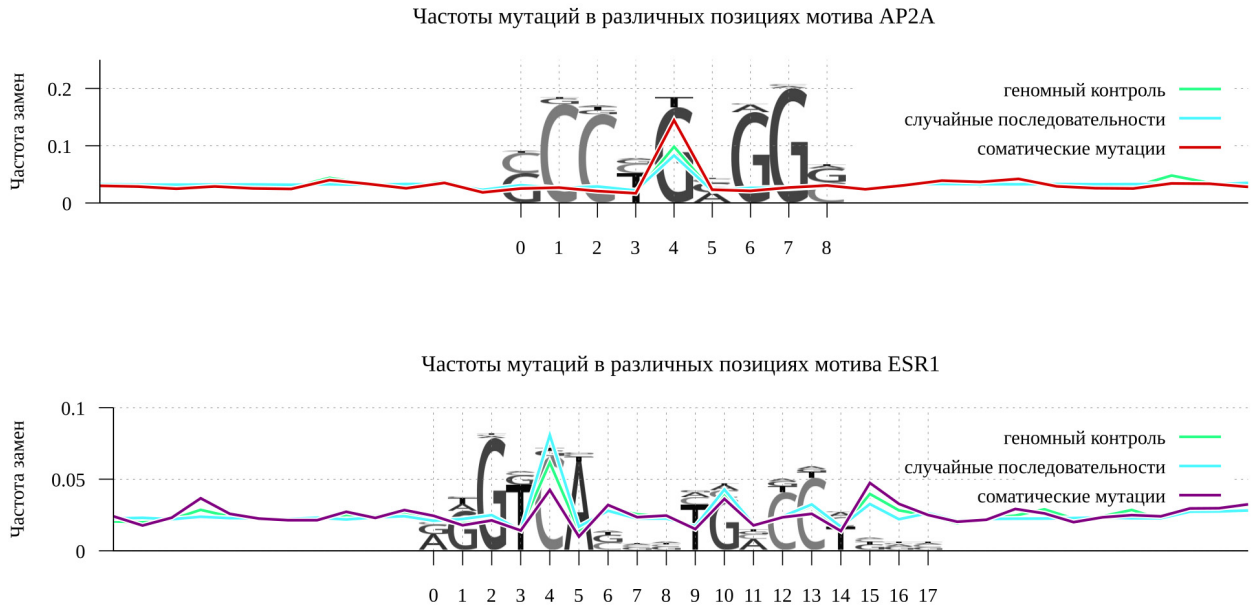


Рисунок 8. Относительное положение мутаций в геноме рака молочной железы по отношению к мотивам AP2A (*верхняя панель*) и ESR1 (*нижняя панель*). Ось Y показывает относительную долю окон, центрированных на мутации, содержащих вхождения мотивов для аллеля зародышевой линии. Ось X соответствует относительному положению мутации. Цвета линий: (*красный*) соматические мутации мотива AP2A, (*фиолетовый*) соматические мутации мотива ESR1, (*голубой и зеленый*) контроли. Рисунок адаптирован из работы [Vorontsov и др., 2016].

Идентификация мотивов в промоторах проекта FANTOM5

Проект FANTOM (Functional ANnoTation Of the Mammalian genome) был начат в институте RIKEN (Институт физико-химических исследований, Япония) в 2000 году для функциональной аннотации неизученных на тот момент полноразмерных кДНК мыши. 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. В ходе проекта FANTOM5 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.

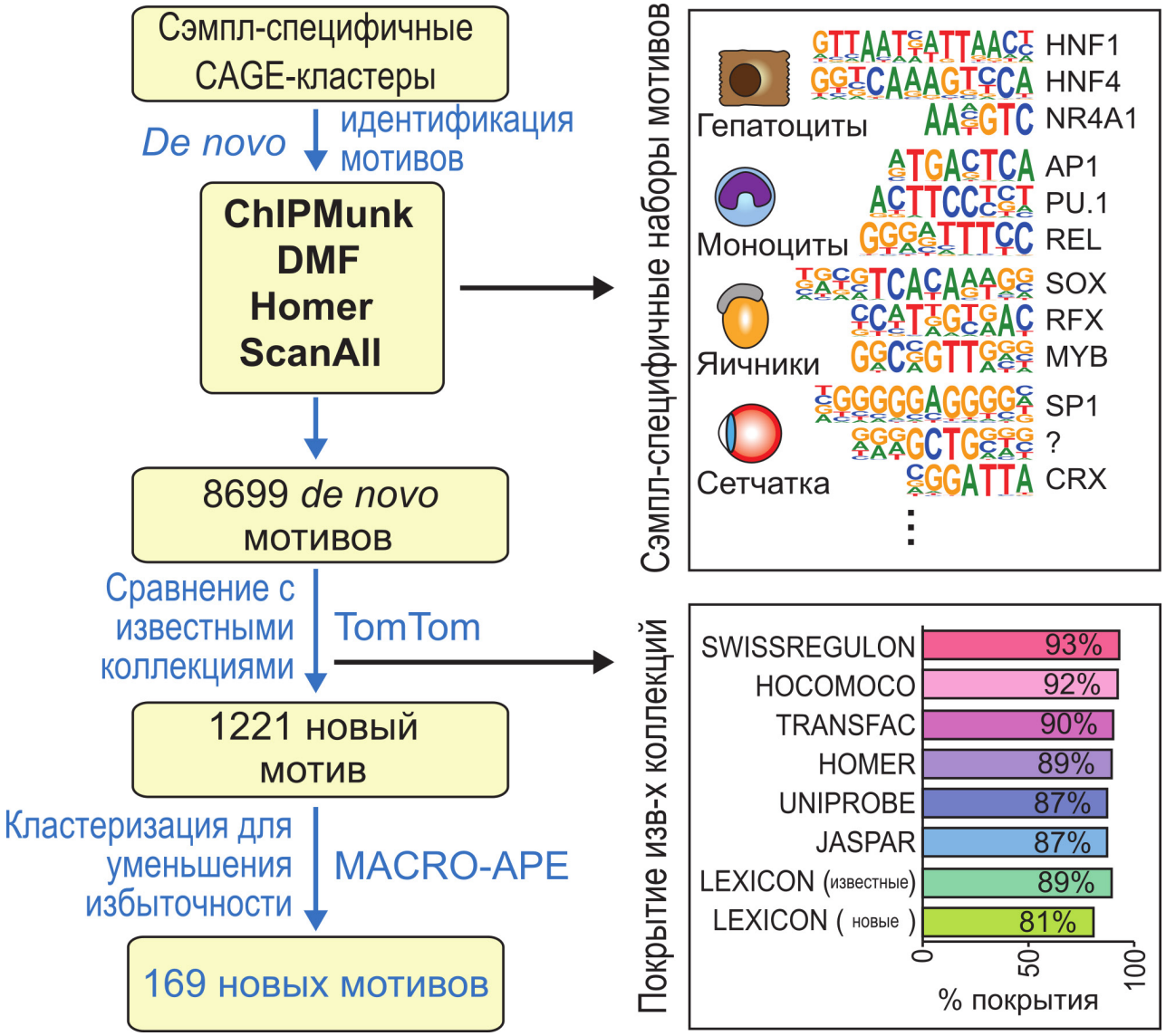


Рисунок 9. Схема идентификации и аннотации мотивов в промоторах, активных в различных биологических образцах (тканях, первичных клетках, иммортализованных клеточных линиях). Рисунок адаптирован из материалов консорциума FANTOM5 [Forrest и др., 2014].

Систематическая идентификация мотивов в ткань-специфичных промоторах была проведена с помощью четырех инструментов: нашего метода ChIPMunk и трех альтернативных методов, использованных членами консорциума, – DMF [Marchand, Bajic, Kaushik, 2011], ScanAll [Zantoni и др., 2007] и HOMER [Heinz и др., 2010]. В сумме это позволило определить более 8 тысяч потенциальных регуляторных мотивов. Важно отметить, что 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 мотивы, характерно отличающиеся от всех известных. Для ответа на эти вопросы использовались актуальные на тот момент релизы коллекций известных мотивов: HOCOMOCO, HOMER, JASPAR; SwissRegulon, UniPROBE и, дополнительно, «регуляторный лексикон» ENCODE, полученный на основе анализа полногеномного футпринтинга ДНКазой I.

С помощью программы TomTom [Gupta и др., 2007] для 7478 из 8699 *de novo*-мотивов были найдены известные аналоги, оставшиеся мотивы потенциально могли соответствовать новым регуляторным сигналам. В обратную сторону, используя коллекцию *de novo*-мотивов как базу данных для поиска, 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 оказались значительно отличались от известных. По сути это означает, что общий репертуар регуляторных мотивов млекопитающих сегодня уже практически изучен.

Процедура, использующая TomTom, позволяет хорошо оценить общую степень похожести наборов мотивов между собой, но 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. С помощью меры Жаккара мы провели иерархическую UPGMA-кластеризацию 1221 «потенциально новых» мотивов, ранее отфильрованных TomTom, и использовали фиксированный порог на длину связи в дереве (соответствующий уровню сходства ~0.05) для получения итогового набора из 172 кластеров. Затем для всех мотивов кластера подсчитывалась корреляция между предсказаниями сайтов связывания и активностью промоторов в различных клетках, мотив с наивысшей корреляцией выбирался как репрезентативный. Схема анализа представлена на **Рисунке 9**. Для 169 кластеров удалось выбрать репрезентативный мотив, вхождения которого были скоррелированы с экспрессией промоторов, и для 37 мотивов наблюдаемая корреляция оказалась значимой с $P < 0.05$ (на основе 1000 случайных перемешиваний колонок весовых матриц с поправкой Бонферрони на множественное тестирование мотивов). Потенциальную функциональную роль новых 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.

Колокализация сайтов связывания факторов транскрипции и CpG-светофоров

Метилирование ДНК является одной из наиболее изученных эпигенетических модификаций, которая выполняет регуляторную функцию в разнообразных нормальных и патологических процессах. В дифференцированных клетках млекопитающих метилирование CG-пар превалирует и является стабильным по отношению к типу клеток, что позволяет говорить о его систематическом вкладе в генную регуляцию. Информация о метилировании отдельных CpG-сайтов обычно используется только для оценки среднего метилирования некоторого района. Вопрос о систематической роли отдельных CpG-сайтов в регуляции транскрипции долгое время оставался открытым, хотя известно, что дифференциальное метилирование конкретных CpG

может влиять на транскрипцию генов [Fürst и др., 2012]. Возможным объяснением ассоциации метилирования отдельных CpG-сайтов с активностью промоторов могла бы быть механистическая роль, например, локализация CpG в сайтах связывания факторов транскрипции и изменение аффинности связывания при метилировании. Так, большинство экспериментально определенных сайтов связывания находятся в неметилированном состоянии [Thurman и др., 2012], и статус метилирования напрямую определяет возможность связывания регулятора [Gebhard и др., 2010]. Возможна и альтернативная модель, в которой метилирование ДНК является не причиной снижения активности гена, а маркером, аккумулируемым при отсутствии связывания факторов транскрипции [Wang и др., 2012]. Чтобы прояснить вопрос о роли метилирования конкретных CpG-сайтов в регуляции экспрессии мы провели совместный систематический анализ позиционных данных по метилированию, ткань-специфичной активности промоторов и сайтов связывания факторов транскрипции.

В данном проекте были использованы данные ENCODE [Dunham и др., 2012] по метилированию ДНК, для оценки активности промоторов в различных типах клеток использовали данные проекта FANTOM5 [Forrest и др., 2014]: 50 образцов ENCODE были сопоставлены с 137 образцами FANTOM5 с усреднением данных для схожих образцов и реплик. Для оценки связи метилирования и экспрессии подсчитывался коэффициент корреляции Спирмана, значимость которого оценивалась с помощью преобразования Фишера. Среди более чем 200 тысяч проанализированных CpG-сайтов, менее процента показали положительную корреляцию экспрессии с метилированием и более 15% (10% для нормальных тканей) – негативную корреляцию ($P < 0.01$). Эти CpG-сайты были названы CpG-светофорами. Почти 80% светофоров локализовались в промоторах генов (-1500 п.н. в 5’ область и до +500 п.н. в 3’ область относительно CAGE-кластера).

Предположим, что CpG-светофоры не являются побочным эффектом от среднего метилирования неактивных промоторов. Тогда отрицательная ассоциация светофоров с экспрессией может объясняться изменением связывания конкретных факторов транскрипции. Следующим шагом был анализ колокализации CpG-светофоров и сайтов связывания факторов транскрипции. Для этого были выполнены полногеномные предсказания сайтов связывания с помощью позиционно-весовых матриц HOCOMOCO v9 (используя пороги оценок по *P*-значениям 0.0005) и моделей RDM с учетом удаленных зависимостей (разработаны в группе проф. В. Байича в KAUST, Саудовская Аравия). Для построения моделей использовались 280 выравниваний с числом сайтов не менее 15, затем учитывались только те факторы транскрипции, для которых удалось пронаблюдать хотя бы одно попадание светофора в предсказанные сайты. В случае RDM, для 100 из 271 рассмотренных факторов транскрипции была обнаружена недопредставленность CpG-светофоров в сайтах связывания (P-значение < 0.05, тест χ^2 с поправкой Бонферрони на множественное тестирование факторов транскрипции). При рассмотрении предсказаний традиционных весовых матриц были получены схожие результаты: сайты 270 из 279 факторов транскрипции избегали перекрытий с CpG-светофорами (**Рисунок 10**).

Изначально 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 данных CTCF, которые также показали систематическое избегание CpG-светофоров. Наконец, мы проанализировали коровые (по порогу на дискретное информационное содержание) 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 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 но, судя по полученным данным, не осуществляется через точечную модификацию аффинности сайтов связывания факторов транскрипции.

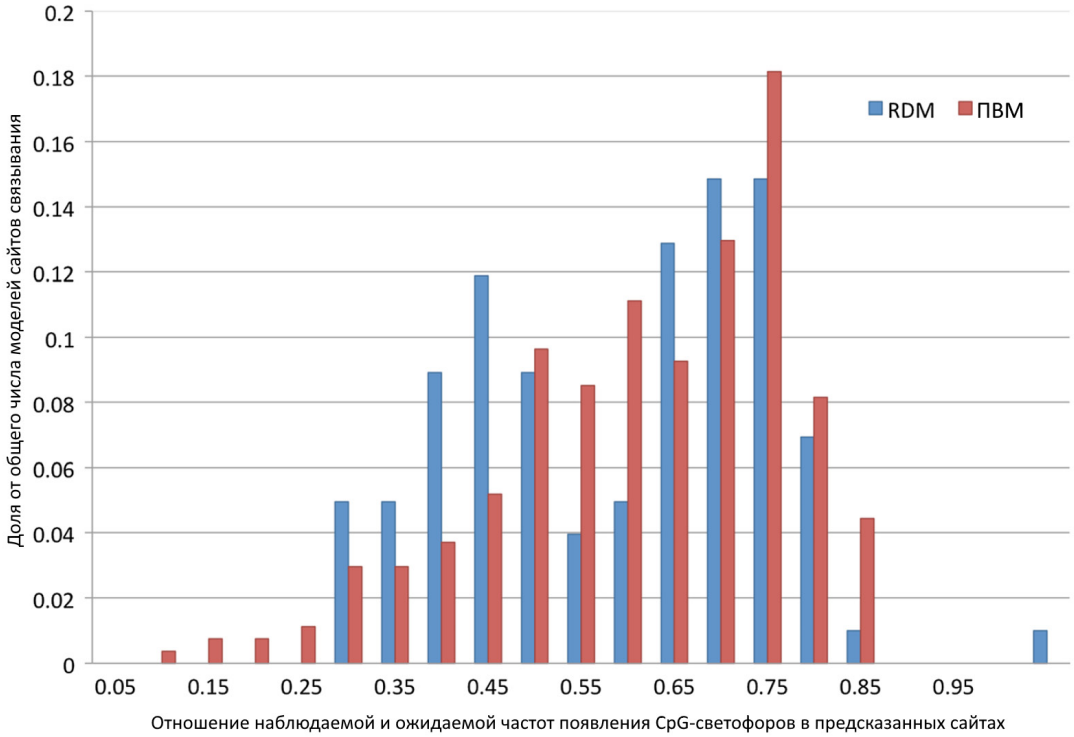


Рисунок 10. Распределение отношений ожидаемых и наблюдаемых частот попадания CpG-светофоров в сайты связывания различных факторов транскрипции. Рисунок адаптирован из работы [Medvedeva и др., 2014].

Заключение

Прочтение индивидуального генома для задач персонализированной медицины уже является решаемой задачей, но все еще невозможна полноценная интерпретация индивидуальных геномных вариантов в некодирующих областях, а ведь именно там находится заметное число казуальных геномных вариантов, например, более 60% вариантов, определяющих предрасположенность к аутоимунным заболеваниям [Farh и др., 2015]. Это обуславливает интерес к функциональной аннотации некодирующих районов генома и стимулирует развитие биоинформатических методов для исследования структуры регуляторных последовательностей. В ходе работы нам удалось создать комплекс 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 у высших эукариот.

Новые биоинформатические методы для анализа мотивов уже способны обрабатывать современные экспериментальные данные большого объема, но на этом развитие не закончено. Для ДНК-паттернов еще не найден наилучший способ представления, пригодный для описания всего спектра экспериментальных данных, а наиболее популярной остается традиционная упрощенная модель мотива – позиционно-весовая матрица. С другой стороны, продолжает дешеветь секвенирование ДНК, но прямой экспериментальный анализ ДНК-белковых комплексов, помимо секвенирования, требует и других серьезных затрат. Это значит, что всеобъемлющий эксперимент для всех факторов транскрипции и сотен типов клеток человеческого организма вряд ли будет проведен в обозримом будущем. Можно надеяться, что для ряда практических приложений альтернативой прямым экспериментальным данным станут достоверные компьютерные предсказания, полученные с помощью современных методов машинного обучения.

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 высших 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.

Публикации по теме диссертации

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

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

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

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