Вавиловский журнал генетики и селекции. 2016;20(6):770-778
DOI 10.18699/VJ16.194

Компьютерный анализ совместной локализации сайтов связывания транскрипционных факторов в геноме по данным ChIP-seq

А.И. Дергилев[1, 2], А.М. Спицина[1], И.В. Чадаева[1, 2], А.В. Свичкарев[1, 3], Ф.М. Науменко[1], Е.В. Кулакова[1], Э.Р. Галиева[1, 2], Е.Е. Витяев[2, 4], М. Чен[5], Ю.Л. Орлов[1, 2]

[1] Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия
[2] Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
[3] Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого», Санкт-Петербург, Россия
[4] Федеральное государственное бюджетное учреждение науки Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук, Новосибирск, Россия
[5] Университет Чжэцзян, Ханчжоу, Китай

Разработана компьютерная программа расчета кластеров сайтов связывания различных транскрипционных факторов (ТФ) по данным геномных координат пиков профиля ChIP-seq (Chromatin ImmunoPrecipitation-sequencing). Рассмотрены статистические особенности распределения сайтов связывания транскрипционных факторов в геноме мыши, полученных с помощью ChIP-seq экспериментов в эмбриональных стволовых клетках. Определены кластеры сайтов, содержащие четыре (и более) сайта связывания различных транскрипционных факторов в геноме мыши, описано их расположение относительно регуляторных районов генов. Подтверждено наличие двух типов совместной локализации сайтов: кластеры, содержащие сайты связывания факторов Oct4, Nanog, Sox2, расположенные в дистальных районах, и кластеры с сайтами связывания n-Myc, c-Myc, находящиеся в основном в промоторных районах генов мыши. Анализ новых данных ChIP-seq по связыванию транскрипционных факторов Nr5a2, Tbx3, Cep, SRF, USF1 в том же типе клеток подтвердил разделение кластеров сайтов связывания транскрипционных факторов на два типа: содержащие сайты связывания регуляторов плюрипотентности (Oct4, Nanog и Sox2) и не включающие их. Разработана компьютерная программа статистической обработки данных о расположении сайтов в генах, использующая экспериментальные данные локализации сайтов, которые получены методами ChIP-seq в геномах мыши и человека. С помощью этой программы выявлены закономерности локализации сайтов связывания транскрипционных факторов различных типов. Рассчитаны расстояния между ближайшими сайтами связывания ТФ группы Oct4, Nanog, Sox2 и сайтами связывания других факторов в кластерах сайтов, которые служат основой для анализа совместного связывания белковых комплексов с ДНК. Рассчитана доля присутствия известных нуклеотидных мотивов сайтов связывания транскрипционных факторов в геномных участках ChIP-seq. Пересчитаны весовые матрицы для таких нуклеотидных мотивов. Показана корреляция присутствия мотивов с интенсивностью связывания ChIP-seq. Программы, реализующие разработанные компьютерные методы оценки кластеризации сайтов связывания различных транскрипционных факторов для новых данных ChIP-seq, доступны по запросу к авторам.

Ключевые слова: сайты связывания; нуклеотидные мотивы; эмбриональные стволовые клетки; поиск закономерностей; ChIP-seq; энхансеры.

Computer analysis of co-localization of transcription factor binding sites in genome by ChIP-seq data

A.I. Dergilev[1, 2], A.M. Spitsina[1], I.V. Chadaeva[1, 2], A.V. Svichkarev[1, 3], F.M. Naumenko[1], E.V. Kulakova[1], E.R. Galieva[1, 2], E.E. Vityaev[2, 4], M. Chen[5], Y.L. Orlov[1, 2]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[3] Peter the Great St. Petersburg Polytechnic University, Saint-Petersburg, Russia
[4] Institute of Mathematics SB RAS, Novosibirsk, Russia
[5] Zhejiang University, Hangzhou, China

Statistical features of the distribution of transcription factor binding sites in the mouse genome that are obtained by ChIP-seq experiments in embryonic stem cells have been considered. Clusters of sites that contain four or more different transcription factor binding sites in the mouse genome have been defined, also their location relatively to the regulatory regions of genes has been described. The presence of two types of site co-localization has been shown: clusters containing binding sites for factors Oct4, Nanog, Sox2, located in the distal regions, and clusters containing binding sites n-Myc, c-Myc, mainly located in the promoter regions of mouse genes. Analysis of new ChIP-seq data about binding of transcription factors Nr5a2, Tbx3 in the same cell type has confirmed the division of clusters of transcription factors binding sites into two types: those containing the binding sites of regulators of pluripotency (Oct4, Nanog, and others) and those not. The computer program of the statistical data processing of gene location and chromatin domains that analyzes experimental data of site localization obtained by ChIP-seq in the mouse genome and the human genome has been developed. The presence of preferences at position of transcription factor binding sites of various types has been revealed, the distances between the nearest groups of TF binding sites Oct4, Nanog, Sox2 and TF binding sites n-Myc and c-Myc have been calculated using this program. The presence

ORIGINAL ARTICLE
Received 15.06.2016
Accepted for publication 30.09.2016

e-mails: arturd1993@yandex.ru; orlov@bionet.nsc.ru

of nucleotide motifs of transcription factor binding sites in the selected areas of ChIP-seq has been estimated, nucleotide motifs have been refined. A correlation between the presence of motifs and the intensity of ChIP-seq binding has been shown. Computer methods for estimating the clustering of different transcription factors binding sites for new data ChIP-seq have been developed. Programs are available upon the request to the authors.

Key words: transcription factor binding sites; embryonic stem cells; data mining; regularity discovery; ChIP-seq; enhancers.

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ:

Дергилев А.И., Спицина А.М., Чадаева И.В., Свичкарев А.В., Науменко Ф.М., Кулакова Е.В., Галиева Э.Р., Витяев Е.Е., Чен М., Орлов Ю.Л. Компьютерный анализ совместной локализации сайтов связывания транскрипционных факторов в геноме по данным ChIP-seq. Вавиловский журнал генетики и селекции. 2016;20(6):770-778. DOI 10.18699/VJ16.194

HOW TO CITE THIS ARTICLE:

Dergilev A.I., Spitsina A.M., Chadaeva I.V., Svichkarev A.V., Naumenko F.M., Kulakova E.V., Galieva E.R., Vityaev E.E., Chen M., Orlov Y.L. Computer analysis of co-localization of transcription factor binding sites in genome by ChIP-seq data. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2016;20(6):770-778. DOI 10.18699/VJ16.194

Исследование регуляции экспрессии генов эукариот в масштабе генома требует изучения сайтов связывания транскрипционных факторов (ССТФ), контролирующих транскрипцию генов, их геномную локализацию, определение генов-мишеней ТФ. В последние годы благодаря методам высокопроизводительного секвенирования ChIP-seq, ChIP-on-chip, ChIP-PET и другим технологиям, сопряженным с иммунопреципитацией хроматина (ChIP – Chromatin ImmunoPrecipitation), появился огромный массив качественно новых данных, позволяющих оценить регуляторный потенциал клетки, в том числе исследовать все сайты связывания заданного транскрипционного фактора в геноме (Орлов, 2014; Игнатьева и др., 2015).

Развитие высокоэффективных экспериментальных методик измерения экспрессии генов, построения карт ДНК-белковых взаимодействий позволяет исследовать совместное расположение сайтов связывания транскрипционных факторов в геноме (Kuznetsov et al., 2007; Chen et al., 2008). Возникают задачи анализа огромных объемов данных и поиска закономерностей организации регуляторных районов генов с помощью статистических, логических и биоинформационных методов (Кулакова и др., 2015; Спицина и др., 2015).

Классическими задачами анализа экспериментальных данных ChIP-seq являются картирование последовательностей прочтений ДНК (ридов), определение пиков геномного профиля связывания, последующая реконструкция нуклеотидных мотивов сайтов связывания по геномным последовательностям, представленным в таких пиках (Орлов, 2014; He et al., 2015). В данной работе мы рассматриваем задачу определения групп (кластеров) сайтов связывания в геноме и анализ расположения мотивов в них.

С появлением первых данных о сайтах связывания ТФ был разработан набор компьютерных методов предсказания сайтов только по нуклеотидной последовательности, поиска закономерностей взаимного расположения сайтов в промоторных районах (Babenko et al., 1999; Витяев и др., 2001; Boeva, 2016). Первые работы основывались на выявлении олигонуклеотидных мотивов и были ограничены статистической базой (десятки, в лучшем случае сотни последовательностей в базах данных) (Heinemeyer et al., 1998). Появление полногеномных методов секвенирования дало толчок к поиску сайтов и регуляторных районов с помощью геномных профилей ChIP-seq, определению предпочтений геномной локализации, нуклеосомной упаковки (Goh et al., 2010; Орлов, 2014).

В этой связи уникальный ресурс по одновременному анализу связывания большого числа различных транскрипционных факторов методами ChIP-seq для анализа как мотивов, так и регуляторных генных сетей представляет работа X. Chen с коллегами (Chen et al., 2008). Эти данные о сайтах связывания послужили основой целого ряда исследований, уточнявших протоколы обработки ChIP-seq, продолжающихся и в настоящее время (Orlov et al., 2009; Kuznetsov et al., 2010; He et al., 2015; Zhang, Wang, 2015). В нашем исследовании рассматривается аспект анализа регуляторных районов на основе кластеризации сайтов связывания различных транскрипционных факторов.

Выбор транскрипционных факторов в упоминаемой выше работе (Chen et al., 2008) был связан с анализом поддержания плюрипотентного состояния клеток. Так, например, транскрипционный фактор Oct4, кодируемый геном *Pou5f1*, – это POU домен-содержащий ТФ, существенный для эмбриональных стволовых клеток (ЭСК) и раннего эмбрионального развития. В частности, Oct4 взаимодействует с Sox2 (HMG-содержащий ТФ), совместно они воздействуют на множество генов в ЭСК человека (Boyer et al., 2005; Loh et al., 2006). В серии работ, включая известную публикацию S. Yamanaka (Takahashi, Yamanaka, 2006), было показано, что Oct4 и Sox2 вместе с c-Myc и Klf4 достаточны для репрограммирования фибробластов в индуцированные плюрипотентные стволовые клетки, которые функционально похожи на ЭСК. Другой транскрипционный фактор – Nanog – представляет собой гомеодомен-содержащий фактор, который может поддерживать состояние плюрипотентности в ЭСК. Однако для этого требуются и другие транскрипционные регуляторы. К ним относятся Esrrb и Zfx, которые регулируют самообновление ЭСК (Ivanova et al., 2006; Loh et al., 2006). Известны также ключевые компоненты сигнальных путей, опосредованных BMP и LIF– Smad1 и STAT3 (Chen et al., 2008). Полногеномный эксперимент ChIP-seq для

фактора Tbx3 в ЭСК мыши показал, что Tbx3 регулирует гены, ассоциированные с состоянием плюрипотентности, и факторы репрограммирования, и, таким образом, относится к группе транскрипционных факторов поддержания плюрипотентности (Han et al., 2010). Еще один важный транскрипционный фактор, E2f1, задействован в регуляции клеточного цикла, где показана ассоциация участков его связывания с промоторными районами генов (Bieda et al., 2006).

Для расширения анализа к данным, полученным в работе X. Chen с коллегами (2008), были добавлены координаты сайтов связывания нескольких транскрипционных факторов, определенные с помощью ChIP-seq в том же типе клеток – Cep, SRF, USF1 (Sirito et al., 1998; Xu et al., 2014; Kuzniewska et al., 2015).

Cep – активатор транскрипции, который связывает ту же последовательность ДНК, что и p53. Он играет важную роль в нормальном развитии, обеспечивает мейозное деление клеток. Следующий транскрипционный фактор, SRF, принадлежит к семейству факторов MADS (MCM1, Agamous, Deficiens и SRF). Этот белок связывается с элементом ответа в сыворотке (SRE) в промоторном участке генов-мишеней. Кроме того, SRF регулирует активность многих ранних генов, например *C-FOS*, и тем самым участвует в регуляции клеточного цикла, апоптоза, клеточного роста и дифференцировки клеток (Kuzniewska et al., 2015). Последний ТФ в нашем добавленном списке – USF1 – кодирует основную спираль лейциновой застежки и может функционировать в качестве фактора клеточной транскрипции (Sirito et al., 1998). Следует отметить, что пониженный уровень USF1 у мышей увеличивает скорость метаболизма.

С помощью компьютерного анализа данных ChIP-seq, доступных в GEONCBI, построены полногеномные карты сайтов связывания транскрипционных факторов в эмбриональных стволовых клетках в геноме мыши для факторов c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2 (Chen et al., 2008; Han et al., 2010; Heng et al., 2010; Lee et al., 2011), а также Cep, SRF, USF1 (Sirito et al., 1998; Xu et al., 2014; Kuzniewska et al., 2015).

В нашей работе были использованы координаты сайтов этих факторов связывания в одном и том же типе клеток, представленные в указанных выше публикациях. Найдена корреляция между присутствием мотивов (процент мотивов в пиках ChIP-seq) и интенсивностью связывания ChIP-seq (высота пика) в геноме мыши. Разработанное программное обеспечение может быть использовано для анализа кластеров сайтов связывания новых наборов транскрипционных факторов в геномах эукариот, включая данные из проектов ENCODE (https://genome.ucsc.edu/ENCODE/), modENCODe, FactorBook (http://www.factorbook.org). Исследование закономерностей расположения нуклеотидных мотивов и контекстных сигналов в геномной ДНК найденных кластеров сайтов с помощью методов интеллектуального анализа данных и знаний (Data Mining) позволит описать точные закономерности структурной организации регуляторных районов генов, в том числе энхансерных районов в геноме для их аннотации и распознавания на основе нуклеотидных последовательностей.

Материалы и методы

В работе были использованы полногеномные карты сайтов связывания транскрипционных факторов в эмбриональных стволовых клетках, построенных по данным ChIP-seq для c-Myc, Oct4, Nanog, Sox2, E2f1, n-Myc, Tbx3, Eset, Nr5a2, Smad2 в геноме мыши (Chen et al., 2008; Han et al., 2010; Heng et al., 2010; Lee et al., 2011). Были изучены закономерности расположения нуклеотидов в кластерах сайтов с помощью программы Match TM и других программ, разработанных на языке Python 2.7 (IDE PyCharm 4.0.4) с использованием библиотек, реализованных на языке C++. Разработан набор утилит командной строки, при этом каждая утилита (на языке Python) реализует одну из требуемых функций. Был задействован ряд скриптов для симуляции числа кластеров в зависимости от общего числа сайтов и размера генома, в том числе только размеров промоторных областей или заданных районов хромосом. С помощью скриптов в среде R были построены тепловые карты и проведены расчеты корреляций встречаемости пар сайтов.

Поиск мотива выполняли с использованием весовой матрицы из базы данных TRANSFAC (Heinemeyer et al., 1998), уточнение – по выравниванию найденных нуклеотидных последовательностей, содержащих мотив в различных наборах данных. Общая схема итеративного компьютерного поиска и уточнения мотивов по данным ChIP-seq представлена на рис. 1.

Результаты и обсуждение

Использованы геномные карты сайтов связывания большого набора транскрипционных факторов в геноме мыши на одном типе клеток (Chen et al., 2008; Han et al., 2010; Heng et al., 2010). Необходимо отметить, что эксперименты иммунопреципитации хроматина могут получать сигнал непрямых ДНК-белковых взаимодействий, поэтому присутствие нуклеотидных мотивов не является необходимым условием и требует дополнительного уточнения. Особенность использованного подхода состоит в изучении одновременного расположения сайтов связывания нескольких различных транскрипционных факторов, предполагается возможность их взаимодействия и согласованного функционирования в исследуемом типе ткани. Показано наличие геномных участков, совместно занятых несколькими сайтами различных факторов, определенных с помощью ChIP-seq. Такие участки можно назвать «горячими точками» совместной локализации ТФ. Они с большой вероятностью представляют собой функционально важные регуляторные участки транскрипции генов. Необходима дальнейшая интеграция данных о связывании сайтов с анализом экспрессии генов (Orlov et al., 2012; Полунин и др., 2014; Спицина и др., 2015) с учетом дальнедействующих взаимодействий хроматина (Li et al., 2014; Кулакова и др., 2015).

Перерасчет карт локализации сайтов связывания Tbx3 (Han et al., 2010) и исследованных ранее кластеров из работы X. Chen с коллегами (2008) показал совместную локализацию связывания Tbx3 с группой Oct4-Sox2-Nanog (Доп. материалы 1)[1].

[1] Дополнительные материалы 1–4 см. в Приложении 1 по адресу: http://www.bionet.nsc.ru/vogis/download/pict-2016-20/appx5.pdf

Статистика расположения сайтов связывания транскрипционных факторов

Анализ расположения сайтов показал, что отдельные участки генома заняты одновременно несколькими различными транскрипционными факторами, связанными с геномной ДНК на очень близком расстоянии (десятки нуклеотидов), или даже с перекрытием участков связывания. Некоторые обогащенные связыванием ТФ районы могут появиться по случайным причинам – близкое расположение сайтов еще не означает их функциональной общности или кооперативного связывания. В то же время некоторые геномные районы, обогащенные сайтами связывания (Chen et al., 2008), могут функционировать как дистальные энхансеры и действительно привлекают кооперативно связывающиеся белковые факторы, физически контактирующие друг с другом при связывании с ДНК. Примеры экспериментального определения таких контактирующих белков, образующих комплекс энхасеосомы, были показаны ранее (Panne et al., 2007) (рис. 2).

Рассмотрим расположение сайтов, определенных с помощью ChIP-seq, в пределах 500 нт. Для статистического разделения неслучайных комбинаций сайтов от «шума» – ожидаемого по случайным причинам числа кластеров сайтов (групп близко расположенных позиций пиков связывания ChIP-seq) – был разработан алгоритм, принимающий во внимание число связанных районов, интенсивность сигнала ChIP-seq в связывании для каждого ТФ. Первый шаг состоял в формальном определении кластера сайтов связывания. Два участка связывания (пики ChIP-seq) включались в кластер, если центральные позиции пиков были удалены не более чем на заданное расстояние (200 нт) друг от друга.

Для оценки вероятности получения таких комбинаций было построено распределение кластеров, которые могут образоваться по случайным причинам, по размерам, в том числе размерам сайтов и хромосом, с использованием подходов, представленных в работах (Chen et al., 2008; Orlov et al., 2009). В целом три сайта связывания различных ТФ в одном и том же геномном локусе могут рассматриваться как неслучайная комбинация. Для более строгого сравнения использовали размер, занимаемый только промоторами генов, без пересечения (определяли 2.5 Кб перед стартом транскрипции и 500 нт – после него), что значительно меньше размера всего генома, доступного для картирования. Связывание четырех и более ТФ одновременно достаточно для принятия гипотезы на уровне 1 % FDR (с вероятностью ошибки ложного предсказания 1 %) как для проксимальных промоторов, так и для дистальных сайтов.

Пики ChIP-seq (сайты связывания) для фиксированного ТФ, содержащие в окрестности 100 нт пики другого фактора, последовательно кластеризовали друг с другом. Кластер увеличивался до тех пор, пока новые пики уже невозможно было добавить. Ограничением по длине геномного участка, варьируемой в программе при построении кластеров, был размер 500 нт. Для каждого локуса получено описание, сколько сайтов разных ТФ, заданных пиками ChIP-seq, он содержит; названия факторов; высота пиков; его геномные координаты (см. рис. 2).

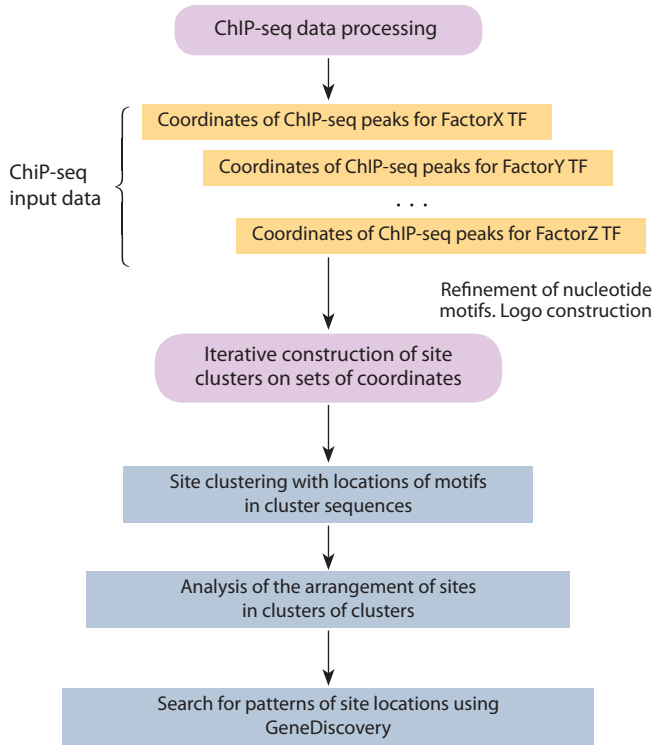

Fig. 1. Data processing: construction of site clusters, search for nucleotide motifs, and analysis of their co-localization.

Рассмотрим геномную локализацию построенных кластеров сайтов связывания транскрипционных факторов. 40 % из них находятся во внутригенных районах, оставшиеся кластеры располагаются в промоторных районах (37 %) и внутри генов (23 %).

Интересно отметить, что кластеры большего размера, как правило, расположены дистально, лишь менее 20 % кластеров из семи и более ТФ находятся в промоторных районах в сравнении с 40 % кластеров размером не более пяти ТФ. Следовательно, совместная встречаемость ССТФ в кластерах не связана с их расположением в промоторах, где могло бы находиться большинство геномных ССТФ. Разработанная программа позволяет пересчитывать уровень значимости для совместной встречаемости группы сайтов для произвольного набора транскрипционных факторов.

Матрицы совместной локализации сайтов связывания различных транскрипционных факторов и тепловые карты

Разработана программа, позволяющая рассчитать по геномным координатам пересечение расположения сайта связывания транскрипционного фактора с другими сайтами и построить кластеры (рис. 3).

На рис. 3 показано число кластеров в зависимости от размера кластера (числа содержащихся в нем сайтов). Параметр расчета границ расположения в данном случае – 200 нт между координатами сайтов, определенными методом ChIP-seq. Из этого рисунка видно, что число сайтов связывания в кластерах в зависимости от размера кластера

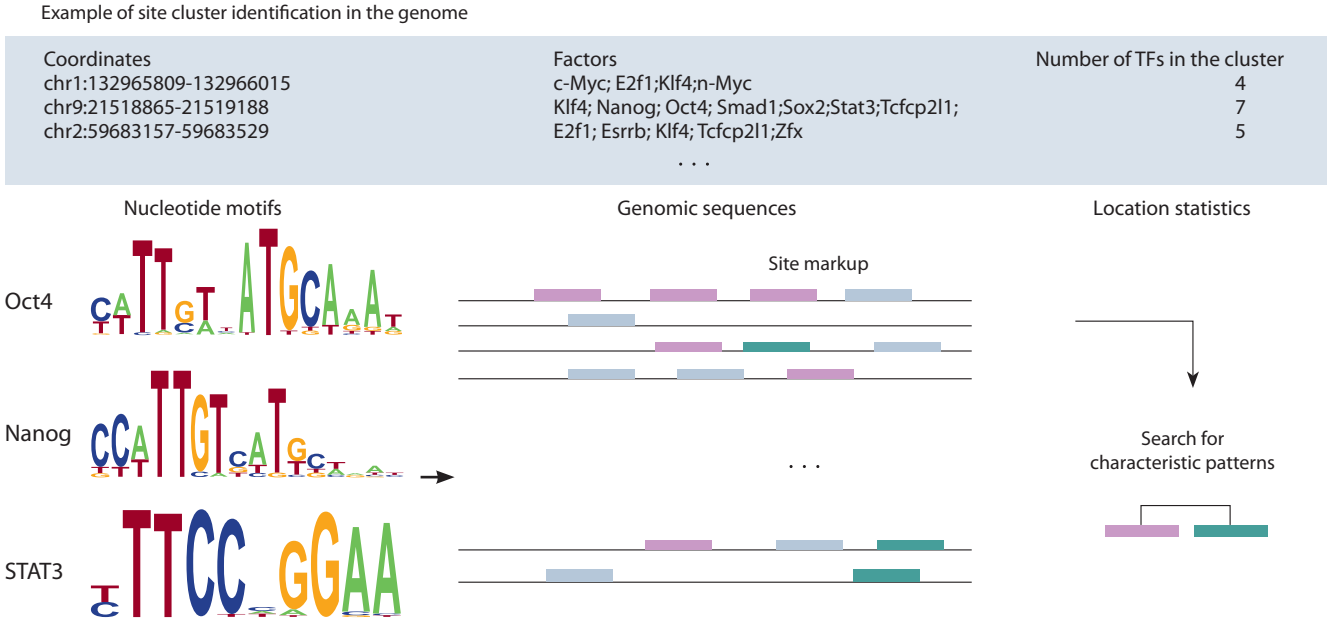

Fig. 2. Schematic presentation of transcription regulation loci containing clusters of transcription factor binding sites recognized according to ChIP-seq data (Chen et al., 2008). Nucleotide motifs of binding sites, their location in regulatory regions, and pattern search.

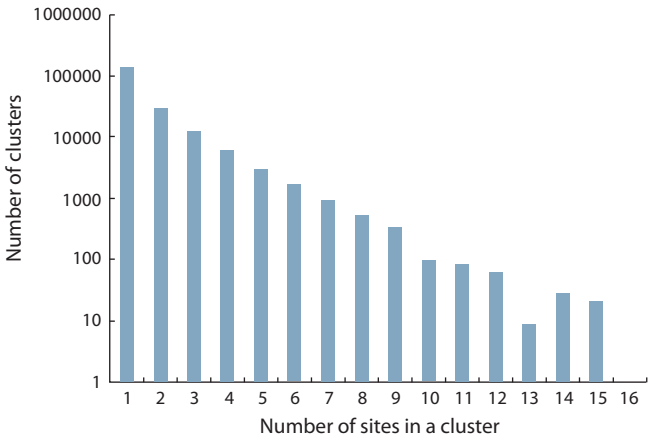

Fig. 3. Distribution of binding site number depending on cluster size.

экспоненциально убывает. Есть единичные кластеры, содержащие одновременно до 15 сайтов связывания различных факторов. В то же время нет геномного района, содержащего одновременно сайты всех 18 факторов, что говорит о функциональных различиях факторов. Подобный результат был получен ранее для набора сайтов связывания 13 факторов, при этом максимальный размер кластера был равен 11 (Orlov et al., 2009) (см. Доп. материалы 2, рис. 1–3).

Для каждой пары транскрипционных факторов был рассчитан коэффициент линейной корреляции Пирсона совместной локализации их сайтов связывания в геноме мыши. Построена матрица частот совместной локализации сайтов различных ТФ по координатам в геноме (рис. 4).

В каждой ячейке такой симметричной матрицы содержится число совпадающих (перекрывающихся в геномном интервале до 200 нт) сайтов связывания двух факторов. По диагонали располагается число сайтов каждого фактора по отдельности. Далее была построена матрица корреляций. Для каждой пары транскрипционных факторов был подсчитан линейный коэффициент корреляции между строками исходной матрицы. Такая матрица задает меру близости, или ассоциации, между различными транскрипционными факторами. Чем выше коэффициент корреляции, тем ближе располагаются сайты связывания транскрипционных факторов друг к другу относительно других факторов. С использованием коэффициента корреляции расположения сайтов как меру близости факторов с помощью программы в среде R была рассчитана матрица совместной локализации исследованных транскрипционных факторов. Интенсивность такой совместной локализации показана в форме тепловой карты (термокарты) (см. рис. 4).

Подтверждено, что группа кластеров Myc (включая гены c-Myc, n-Myc) характеризуется преимущественно промоторным расположением, а группа Nanog (с другими ключевыми факторами плюрипотентности) – преимущественно дистальным расположением относительно старта транскрипции генов.

Выявление нуклеотидных мотивов в кластерах

Анализ данных ChIP-seq о сайтах связывания нескольких транскрипционных факторов позволяет получить комбинаторную информацию о регуляторном действии сайтов (Chen et al., 2008). Такая информация собирается с помощью определения точного положения нуклеотидных мотивов и их взаимного расположения в регуляторных участках генов (в нашем случае кластеров сайтов).

Проведена обработка исходных данных ChIP-seq, рассчитаны пороги для выделения пиков из профиля и число пиков и выделены их координаты в геноме мыши. Результаты представлены в таблице, в которой показано число сайтов для каждого транскрипционного фактора по отдельности, варьирующее от 1 до 39 тыс. сайтов в геноме. Длина нуклеотидного мотива (весовая матрица) взята из базы данных TRANSFAC (Heinemeyer et al., 1998). Далее был рассчитан процент присутствия нуклеотидных мотивов в пиках ChIP-seq всех анализируемых ТФ.

Процент присутствия вычисляли как отношение числа пиков с хотя бы одним присутствием мотива (выше или равно порога по score, порог был установлен равным 0.8) к общему числу пиков данного ТФ. Процент присутствия мотивов характеризует как информационное содержание мотива, так и качество данных ChIP-seq в наборе. При этом присутствие или отсутствие выраженного нуклеотидного мотива недостаточно для определения качества эксперимента, как показано в ряде работ на том же наборе данных (Kuznetsov et al., 2010). Для пиков ChIP-seq в кластерах, содержащих несколько различных сайтов, процент содержания нуклеотидных мотивов выше, чем для пиков ChIP-seq, расположенных вне кластеров. Таким образом, присутствие других сайтов (пиков ChIP-seq) на близком расстоянии в геноме может свидетельствовать о большей функциональной значимости рассматриваемого регуляторного участка. Дальнейший анализ олигонуклеотидных мотивов в сайтах (Putta et al., 2011) может служить контекстной характеристикой для изучения регуляторных районов транскрипции генов в целом.

Рассмотрим пример уточнения нуклеотидного мотива для сайтов связывания c-Myc. (Полностью данные для набора всех ТФ приведены в Доп. материалах 3.) Пример уточнения нуклеотидных мотивов по сравнению с известными ранее в базах данных (TRANSFAC и JASPAR) показан на рис. 5.

В Доп. материалах 3 (таблица) показано графическое представление мотивов для всех рассмотренных ТФ.

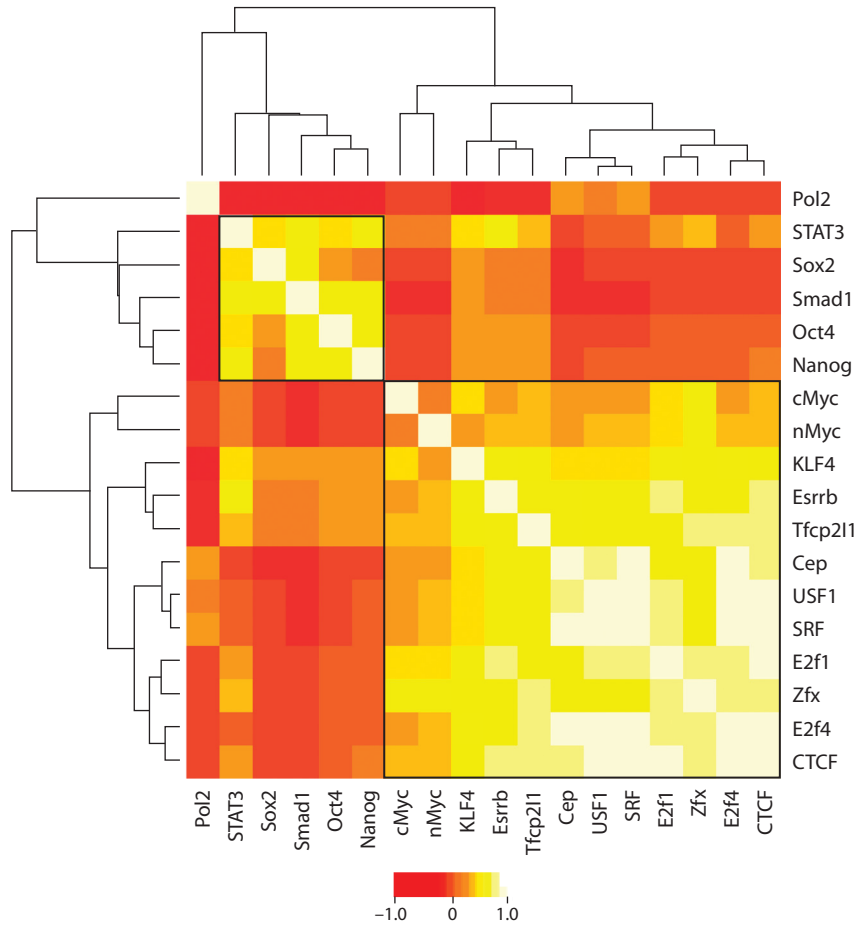

Fig. 4. Heatmap of the colocalization of transcription factor binding sites in the mouse genome according to ChIP-seq data.

Calculation of the colocalization of sites with regard to new factors was performed as in (Chen et al., 2008).

Для сайтов связывания CTCF показана высокая консервативность мотива (Доп. материалы 4, рис. 1). Выполнена качественная оценка присутствия мотивов в зависимости от интенсивности связывания (высоты пиков). Пики ChIP-seq были сортированы по высоте (интенсивности связывания), полученный ранжированный список разбит на квартили. В каждой квартили подсчитаны число и относительная доля найденных мотивов. Показана положительная ассоциация присутствия мотивов в зависимости от квартили (25 % элементов упорядоченного списка). Результаты приведены в Доп. материалах 4, на рис. 2 и 3. Таким образом, анализ нуклеотидных мотивов позволяет повысить качество рассматриваемых данных для дальнейшего исследования при выборе только пиков ChIP-seq, содержащих мотивы.

Были проанализированы состав и предпочтения к совместной локализации транскрипционных факторов в кластерах сайтов. В целом в исследованном наборе выделяются две группы: относящиеся к Myc и относящиеся к Nanog транскрипционные факторы. Среди 18 факторов Nanog, Sox2, Oct4, Smad1, и STAT3 имеют тенденцию встречаться совместно чаще (см. рис. 4). Сайты связывания Zfx, CTCF и E2f1 чаще встречаются совместно в промоторных районах генов, так же, как и кластеры сайтов, содержащие сайты Myc. Недавние исследования (Yanan et al., 2016) показали, что связывание активного CTCF (CCCTC-связывающего фактора), который имеет важное значение для регуляции генов, с CBS (CTCF-сайтом связывания) устанавливает специфическое взаимодействие между энхансерами и промоторами, а три кластера Pcdhβγ образуют два CCD (CTCF/ когезин-опосредованных домена хрома-

Binding site motifs in ChIP-seq peaks

Transcription factor	Total number of ChIP-seq peaks	Nucleotide motif length	Percentage of present motifs
c-Myc	3422	19	58.82
CTCF	39609	8	75.53
E2f1	20699	12	99.90
E2f4	4546	10	60.70
Esrrb	21647	16	83.57
KLF4	10875	10	27.58
Nanog	10343	15	58.25
n-Myc	7182	16	44.45
Oct4	3761	15	24.42
Pol2	30877	10	38.57
Smad1	1126	14	59.42
Sox2	4526	12	46.02
SRF	2609	10	43.04
Cep	11434	10	–
STAT3	2546	10	–
Tfcp2l1	26910	10	–
USF1	6741	10	–
Zfx	10368	10	–

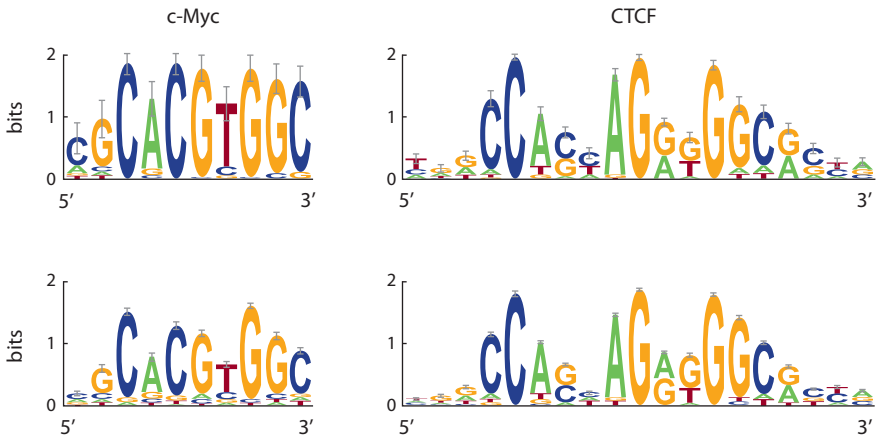

Fig. 5. Known nucleotide motifs (logos) of binding sites of transcription factors c-Myc and CTCF.

Above: known motifs of c-Myc and CTCF; below: motifs refined according to nucleotide sequences from ChIP-seq peaks.

тина). Полагается, что CTCF опосредуют специфические взаимодействия между энхансерами с образованием активной площадки транскрипции для экспрессии генов. Кроме того, в ходе исследований (Hutter et al., 2010) было обнаружено умеренное обогащение сайтов связывания фактора CTCF в области геномного импринтинга. Было найдено несколько мотивов в элементах с высокой консервативностью, которые могут выступать в качестве дополнительных регуляторных элементов.

Заключение

Анализ расширенного набора сайтов связывания транскрипционных факторов подтвердил совместную кластеризацию тех сайтов, которые относятся к поддержанию плюрипотентности – Nanog, Sox2, Oct4 – относительно более широкого набора факторов, координаты которых определены экспериментально в статье X. Chen с коллегами (2008). Объединение полногеномных данных картирования сайтов с помощью программ анализа данных ChIP-seq позволяет исследовать генные сети регуляции плюрипотентного состояния стволовых клеток и в дальнейшем рассматривать возможности оптимизации репрограммирования клеток (Orlov et al., 2012).

Использование нуклеотидных последовательностей кластеров сайтов позволяет уточнить нуклеотидные мотивы связывания. С помощью компьютерного исследования комбинаций сайтов связывания по данным ChIP-seq можно определять более сложные закономерности с использованием дополнительных характе-

ристик нуклеотидных последовательностей, в частности оценки сложности текста (Babenko et al., 2015; Safronova et al., 2015), и их функциональную аннотацию, полученную с помощью других пакетов, таких как ICGenomics (Орлов и др., 2012).

Анализ расположения кластеров относительно генов с помощью таких инструментов, как GREAT (Genomic Regions Enrichment of Annotations Tool), дает возможность оценить расположение сайтов связывания отдельных факторов и кластеров сайтов относительно генов (Guo et al., 2012), классифицировать их как промоторные и дистальные, что является дополнительной функциональной характеристикой таких кластеров. Дополнительные характеристики нуклеотидных последовательностей для описания кластеров сайтов связывания, таких как участки низкой сложности текста (Orlov, Potapov, 2004; Orlov et al., 2006), нуклеосомная упаковка (Орлов и др., 2006; Goh et al., 2010), позволяют более точно определить регуляторные районы в геноме и выполнять их поиск на основе непрямых данных, без экспериментов ChIP-seq. Интеграция геномных данных средствами UGENE (Vas'kin et al., 2011; Васькин и др., 2012; Golosova et al., 2014) позволит качественно решать новые задачи комбинаторного анализа сайтов по данным проектов ENCODE (https://genome.ucsc.edu/ENCODE/) и FactorBook (http://www.factorbook.org) в геноме человека. Развитие методов поиска закономерностей взаимного расположения сайтов связывания в регуляторных районах (комбинаций сайтов ChIP-seq в кластерах) по уточненным мотивам способствует определению более тонких закономерностей работы транскрипционных факторов для регуляции транскрипции генов (Boeva, 2016).

Acknowledgments

We are grateful to Drs. Yu.Yu. Vas'kin, K.S. Bekker, N.S. Safronova, and M.S. Evdokimov for fruitful discussion. Computation was done at the Bioinformatics Shared Access Center, SB RAS, and the Siberian Supercomputer Center, SB RAS. The study of transcription factor binding sites was supported by the Russian Foundation for Basic Research, project 14-04-01906. The development of programs for genomic data analysis was supported by the Russian Science Foundation, project 14-24-00123.

Conflict of interest

The authors declare no conflict of interest.

References

Babenko V.N., Kosarev P.S., Vishnevsky O.V., Levitsky V.G., Basin V.V., Frolov A.S. Investigating extended regulatory regions of genomic DNA sequences. Bioinformatics. 1999;15(7-8):644-653. DOI 10.1093/bioinformatics/15.7.644.

Babenko V.N., Matvienko V.F., Safronova N.S. 19 Implication of transposons distribution on chromatin state and genome architecture in human. J. Biomol. Struct. Dyn. 2015;33(1):10-11. DOI 10.1080/07391102.2015.1032559.

Bieda M., Xu X., Singer M.A., Green R., Farnham P. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res. 2006;16(5):595-605. DOI 10.1101/gr.4887606.

Boeva V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. Front. Genet. 2016;7:24. DOI 10.3389/fgene.2016.00024.

Boyer L.A., Lee T.I., Cole M.F., Johnstone S.E., Levine S.S., Zucker J.P., Guenther M.G., Kumar R.M., Murray H.L., Jenner R.G., Gifford D.K., Melton D.A., Jaenisch R., Young R.A. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005;122(6):947-956. DOI 10.1016/j.cell.2005.08.020.

Chen X., Xu H., Yuan P., Fang F., Huss M., Vega V.B., Wong E., Orlov Y.L., Zhang W., Jiang J., Loh Y.H., Yeo H.C., Yeo Z.X., Narang V., Govindarajan K.R., Leong B., Shahab A., Ruan Y., Bourque G., Sung W.K., Clarke N.D., Wei C.L., Ng H.H. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell. 2008;133(6):1106-1117. DOI 10.1016/j.cell.2008.04.043.

Goh W.S., Orlov Y., Li J., Clarke N.D. Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. PLoS Comput. Biol. 2010;6(1):e1000649. DOI 10.1371/journal.pcbi.1000649.

Golosova O., Henderson R., Vas'kin Yu., Gabrielian A., Grekhov G., Nagarajan V., Oler A.J., Quiñones M., Hurt D., Fursov M., Huyen Y. Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses. Peer J. 2014;2:e644. DOI 10.7717/peerj.644.

Guo Y., Mahony S., Gifford D.K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Comput. Biol. 2012;8(8):e1002638. DOI 10.1371/journal.pcbi.1002638.

Han J., Yuan P., Yang H., Zhang J., Soh B.S., Li P., Lim S.L., Cao S., Tay J., Orlov Y.L., Lufkin T., Ng H.H., Tam W.L., Lim B. Tbx3 improves the germ-line competency of induced pluripotent stem cells. Nature. 2010;463(7284):1096-1100.

He X., Cicek A.E., Wang Y., Schulz M.H., Le H.-S., Ziv B.-J. De novo ChIP-seq analysis. Genome Biol. 2015;16(1):205. DOI 10.1186/s13059-015-0756-4.

Heinemeyer T., Wingender E., Reuter I., Hermjakob H., Kel A.E., Kel O.V., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Kolpakov F.A., Podkolodny N.L., Kolchanov N.A. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. Nucleic Acids Res. 1998;26(1):362-367. DOI 10.1093/nar/26.1.362.

Heng J.C., Feng B., Han J., Jiang J., Kraus P., Ng J.H., Orlov Y.L., Huss M., Yang L., Lufkin T., Lim B., Ng H.H. The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. Cell Stem Cell. 2010;6(2):167-174. DOI 10.1016/j.stem.2009.12.009.

Hutter B., Bieg M., Helms V., Paulsen M. Imprinted genes show unique patterns of sequence conservation. BMC Genomics. 2010;11:649. DOI 10.1186/1471-2164-11-649.

Ignatieva E.V., Podkolodnaya O.A., Orlov Yu.L., Vasiliev G.V., Kolchanov N.A. Regulatory genomics: Combined experimental and computational approaches. Rus. J. Genet. 2015;51(4):334-352. DOI 10.1134/S1022795415040067.

Ivanova N., Dobrin R., Lu R., Kotenko I., Levorse J., DeCoste C., Schafer X., Lun Yi., Lemischka I.R. Dissecting self-renewal in stem cells with RNA interference. Nature. 2006;442(7102):533-538. DOI 10.1038/nature04915.

Kuznetsov V.A., Orlov Yu.L., Wei C.L., Ruan Y. Computational analysis and modeling of genome-scale avidity distribution of transcription factor binding sites in chip-pet experiments. Genome Inform. 2007;19:83-94.

Kulakova E.V., Spitsina A.M., Orlova N.G., Dergilev A.I., Svichkarev A.V., Safronova N.S., Chernykh I.G., Orlov Yu.L. Programs for the analysis of genomic sequence data obtained by ChIP-seq, ChIA-PET, and Hi-C technologies. Programmnye Sistemy: Teoriya i Prilozheniya = Program Systems: Theory and Applications. 2015;6(2(25)):129-148. (in Russian)

Kuznetsov V.A., Singh O., Jenjaroenpun P. Statistics of protein-DNA binding and the total number of binding sites for a transcription factor in the mammalian genome. BMC Genomics. 2010;11(1):S12. DOI 10.1186/1471-2164-11-S1-S12.

Kuzniewska B., Nader K., Dabrowski M., Kaczmarek L., Kalita K. Adult deletion of SRF increases epileptogenesis and decreases ac-

tivity-induced gene expression. Mol. Neurobiol. 2015;1-16. DOI 10.1007/s12035-014-9089-7.

Lee K.L., Lim S.K., Orlov Y.L., Yit le Y., Yang H., Ang L.T., Poellinger L., Lim B. Graded Nodal/Activin signaling titrates conversion of quantitative phospho-Smad2 levels into qualitative embryonic stem cell fate decisions. PLoS Genet. 2011;7(6):e1002130. DOI 10.1371/journal.pgen.1002130.

Li G., Cai L., Chang H., Hong P., Zhou Q., Kulakova E.V., Kolchanov N.A., Ruan Y. Chromatin interaction analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. BMC Genomics. 2014;15(12):S11. DOI 10.1186/1471-2164-15-S12-S11.

Loh Y.H., Wu Q., Chew J.L., Vega V.B., Zhang W., Chen X., Bourque G., George J., Leong B., Liu J., Wong K.Y., Sung K.W., Lee C.W., Zhao X.D., Chiu K.P., Lipovich L., Kuznetsov V.A., Robson P., Stanton L.W., Wei C.L., Ruan Y., Lim B., Ng H.H. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat. Genet. 2006;38(4):431-440. DOI 10.1038/ng1760.

Orlov Yu.L. Computer-assisted study of the regulation of eukaryotic gene transcription on the base of data on chromatin sequencing and precipitation. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2014;18(1):193-206. (in Russian)

Orlov Yu.L., Bragin A.O., Medvedeva I.V., Gunbin K.V., Demenkov P.S., Vishnevsky O.V., Levitsky V.G., Oshchepkov D.Y., Podkolodnyy N.L., Afonnikov D.A., Grosse I., Kolchanov N.A. IC-Genomics: Software for analysis of symbol genomics sequences. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2012;16(4/1):732-741. (in Russian)

Orlov Yu.L., Huss M.E., Joseph R., Xu H., Vega V.B., Lee Y.K., Goh W.S., Thomsen J.S., Cheung E.C., Clarke N.D., Ng H.H. Genome-wide statistical analysis of multiple transcription factor binding sites obtained by ChIP-seq technologies. Proc. 1st ACM Workshop on Breaking Frontiers of Computational Biology (CompBio '09). ACM, New York. N.Y., 2009;11-18.

Orlov Yu.L., Levitskii V.G., Smirnova O.G., Podkolodnaya O.A., Khlebodarova T.M., Kolchanov N.A. Statistical analysis of nucleosome formation sites. Biofizika = Biophysics. 2006;51(4):608-614. (in Russian)

Orlov Yu.L., Potapov V.N. Complexity: an internet resource for analysis of DNA sequence complexity. Nucleic Acids Res. 2004;32:W628-W633. DOI 10.1093/nar/gkh466.

Orlov Yu.L., Te Boekhorst R., Abnizova I.I. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. J. Bioinform. Comput. Biol. 2006;4:523-536. DOI 10.1142/S0219720006001801.

Orlov Yu., Xu H., Afonnikov D., Lim B., Heng J.C., Yuan P., Chen M., Yan J., Clarke N., Orlova N., Huss M., Gunbin K., Podkolodnyy N., Ng H.H. Computer and statistical analysis of transcription factor binding and chromatin modifications by ChIP-seq data in embryonic stem cell. J. Integr. Bioinform. 2012;9(2):211. DOI 10.2390/biecoll-jib-2012-211.

Panne D., Maniatis T., Harrison S.C. An atomic model of the interferon-beta enhanceosome. Cell. 2007;129(6):1111-1123. DOI 10.1016/j.cell.2007.05.019.

Polunin D.A., Shtaiger I.A., Efimov V.M. JACOBI 4 software for multivariate analysis of microarray data. Vestnik NGU. Ser. Informatsionnye tekhnologii = Novosibirsk State University Journal of Information Technologies. 2014;12(2):90-98. (in Russian)

Putta P., Orlov Yu.L., Podkolodnyy N.L., Mitra C.K. Relatively conserved common short sequences in transcription factor binding sites and miRNA. Vavilov Journal Genetics and Breeding. 2011;15(4):750-756.

Safronova N.S., Babenko V.N., Orlov Yu.L. 117 Analysis of SNP containing sites in human genome using text complexity estimates. J. Biomol. Struct. Dyn. 2015;33(1):73-74. DOI 10.1080/07391102.2015.1032750.

Sirito M., Lin Q., Deng J.M., Behringer R.R., Sawadogo M. Overlapping roles and asymmetrical cross-regulation of the USF proteins in mice. Overlapping roles and asymmetrical cross-regulation of the USF proteins in mice. Proc. Natl. Acad. Sci. USA. 1998;95(7):3758-3763.

Spitsina A.M., Orlov Yu.L., Podkolodnaya N.N., Svichkarev A.V., Dergilev A.I., Chen M., Kuchin N.V., Chernych I.G., Glinskiy B.M. Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing. Programmnye Sistemy: Teoriya i Prilozheniya = Program Systems: Theory and Applications. 2015;6(1(23)):157-174. (in Russian)

Takahashi K., Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006;126(4):663-676. DOI 10.1016/j.cell.2006.07.024.

Vas'kin Yu., Khomicheva I.V., Ignatieva E.V., Vityaev E.E. Expert discovery and UGENE integrated system for intelligent analysis of regulatory regions of genes. In Silico Biol. 2011-2012;11(3-4):97-108. DOI 10.3233/ISB-2012-0448.

Vas'kin Yu.Yu., Khomicheva I.V., Ignatieva E.V., Vityaev E.E. Sequence analysis of regulatory regions of genes with the Expert Discovery relational system built in package UGENE. Vestnik NGU. Ser. Informatsionnye tekhnologii = Novosibirsk State University Journal of Information Technologies. 2012;10(1):73-86. (in Russian)

Vityaev E.E. Izvlechenie znaniy iz dannykh. Kompyuternoe poznanie. Modeli kognitivnykh protsessov [Data mining: Computerized cognition. Models of cognitive processes]. Novosibirsk: Novosibirsk State University Publ., 2006. (in Russian)

Vityaev E.E., Orlov Yu.L., Vishnevsky O.V., Belenok A.S., Kolchanov N.A. Computer system "Gene Discovery" to search for patterns in eukaryotic regulatory nucleotide sequences. Molekulyarnaya biologiya = Molecular Biology (Moscow). 2001;35(6):952-960. (in Russian)

Xu D., Wei G., Lu P., Luo J., Chen X., Skogerb G., Chen R. Analysis of the p53/CEP-1 regulated non-coding transcriptome in *C. elegans* by an NSR-seq strategy. Protein Cell. 2014;5(10):770-782. DOI 10.1007/s13238-014-0071-y.

Yanan Z., Quan X., Ya G., Qiang W. Characterization of a cluster of CTCF-binding sites in a protocadherin regulatory region. Yi Chuan. 2016;38(4):323-336. DOI 10.16288/j.yczz.16-037.

Zhang Y., Wang P. A fast cluster motif finding algorithm for ChIP-Seq data sets. Biomed. Res. Int. 2015;2015;218068. DOI 10.1155/2015/218068.