

А. В. Цуканов^{1,2}, Н. Г. Орлова³, А. И. Дергилев², Ю. Л. Орлов^{1,2}

¹ *Институт цитологии и генетики СО РАН
пр. Академика Лаврентьева, 10, Новосибирск, 630090, Россия*

² *Новосибирский государственный университет
ул. Пирогова, 1, Новосибирск, 630090, Россия*

³ *Новосибирский государственный архитектурно-строительный университет (Сибстрин)
ул. Ленинградская, 113, Новосибирск, 630008, Россия*

ya.cukanton@yandex.ru, orlovanina2@mail.ru, arturd1993@yandex.ru, orlov@bionet.nsc.ru

ПРОГРАММЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ, КЛАСТЕРИЗАЦИИ И ВИЗУАЛИЗАЦИИ РАСПРЕДЕЛЕНИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ В ГЕНОМЕ *

Исследование регуляции транскрипции генов на основе данных современных технологий высокопроизводительного секвенирования является актуальной задачей биоинформатики, требующей развития новых компьютерных средств, в том числе на основе суперкомпьютерных вычислений. Рассмотрены задачи обработки данных полногеномных профилей ChIP-seq связывания транскрипционных факторов в геномах, определения пиков профилей и поиска сайтов связывания в нуклеотидных последовательностях таких пиков. Разработаны программы для анализа положения сайтов связывания в геноме относительно районов генов, расчета кластеров таких сайтов и визуализации их положения в геноме. Рассчитаны кластеры сайтов связывания транскрипционных факторов в геноме человека по базе данных Cistrome, построены матрицы совместной встречаемости пар сайтов связывания различных транскрипционных факторов в геноме для различных типов тканей и культур клеток. Проведен вычислительный эксперимент по компьютерной генерации случайных кластеров в геноме, а также оценке встречаемости кластеров большого размера для экспериментально полученных сайтов связывания транскрипционных факторов в геноме человека. Найдены закономерности встречаемости сайтов факторов плюрипотентности в эмбриональных стволовых клетках. Разработанное программное обеспечение доступно по запросу к авторам.

Ключевые слова: геномика, секвенирование, сайты связывания транскрипционных факторов, промоторы, большие данные, статистика, визуализация.

Введение

Изучение структурно-функциональной организации генома на основе данных высокопроизводительного секвенирования ДНК продолжает оставаться магистральным направлением, развивающимся на стыке биологии и информационных технологий. Исследование организации генетической информации в геноме (на различных уровнях - ДНК, РНК и белки) требует применения современных технологических подходов высокопроизводительного секвенирования, что, в свою очередь, определяет биоинформационные задачи: первичная обработка сырых данных (картирование прочтений ДНК), анализ массивов данных (анализ дифферен-

* Работа была поддержана РФФИ и бюджетным проектом ИЦиГ СО РАН (№ 0324-2018-0017).

Авторы благодарны Ирине Вадимовне Медведевой, Владимиру Николаевичу Бабенко и Антону Геннадьевичу Богомолу за помощь в работе, и научную дискуссию.

циальной экспрессии генов, анализ пиков профилей ChIP-seq) и аннотация генома на основе таких прочтений (аннотирование транскриптов, определение положения сайтов связывания транскрипционных факторов) [1].

Ключевую роль в работе клетки играет регуляция транскрипции генов, которые определяют свойства клетки [2]. Изучение проблемы регуляции, развитие методов и технологий секвенирования привели к накоплению огромного количества данных, полученных различными методами секвенирования, такими как Hi-C, ChIP-seq, BS-seq, DNaseI-seq, ATAC-seq, NOMe-seq, RNA-seq [1; 3; 4]. Доступны большие объемы данных как для генома человека, так и для геномов других модельных организмов эукариот (мышь, крыса, растения). Рассмотрим несколько уровней регуляции транскрипции с точки зрения анализа данных [4].

Первый и, возможно, основной уровень – это регуляция на уровне ДНК, здесь регуляция осуществляется за счет связывания транскрипционных факторов с сайтами на ДНК, что может приводить как к активации транскрипции, так и к ее остановке [5]. Также на этом уровне выделяется влияние метилирования ДНК, которое может оказывать влияние на регуляцию за счет метилирования сайтов связывания транскрипционных факторов (ССТФ), что приводит к изменению сродства транскрипционного фактора (ТФ) к его сайту [6].

Следующий уровень – это регуляция на уровне хроматина, где в зависимости от модификаций гистонов может меняться структура хроматина, он может быть в открытом или закрытом состоянии, что влияет на транскрипцию генов. Состояние хроматина определяется по данным анализа геномных профилей секвенирования. В закрытом состоянии хроматина транскрипционные факторы и транскрипционная «машина» эукариотической клетки неспособна подобраться к сайту начала транскрипции из-за физической недоступности ДНК для этих белков [7; 8].

Третий немаловажный уровень регуляции осуществляется за счет трехмерной структуры хромосом. Трехмерная упаковка в ядре клетки может оказывать значительное влияние на регуляцию экспрессии генов за счет сближения удаленных регуляторных участков (энхансеров, сайленсеров) и промоторов генов, тем самым регулируя уровень экспрессии генов [9; 10].

Несмотря на большой массив накопленной информации, остаются пробелы в знаниях о регуляции экспрессии генов, что приводит к необходимости продолжать исследования, связанные с анализом регуляции экспрессии генов эукариот в масштабе генома, и разработку соответствующих программных инструментов. Исследование регуляции экспрессии генов эукариот в масштабе генома требует изучения сайтов связывания транскрипционных факторов (ССТФ), контролирующих транскрипцию генов, их геномной локализации, определения их генов-мишеней [3; 11]. Благодаря развитию методов высокопроизводительного секвенирования ChIP-seq, ChIP-on-chip и другим технологиям, сопряженным с иммунопреципитацией хроматина (ChIP – Chromatin ImmunoPrecipitation) [4], стал доступен огромный массив новых данных, позволяющих исследовать все сайты связывания заданного транскрипционного фактора в геноме, а также комбинации таких сайтов [3; 12].

Экспериментально установленное число сайтов в геноме может варьировать от нескольких сотен до десятков тысяч [1; 3; 4; 12]. Значительная часть ССТФ располагается в дистальных (удаленных) районах генов, что затрудняет точное определение тех генов, транскрипцию которых они регулируют. Встают задачи анализа регуляторных районов генов, в том числе дистальных, поиска закономерностей расположения в них сайтов и контекстных сигналов с помощью статистических, логических и биоинформационных методов [11; 13; 14].

Исследование влияния близко расположенных и пересекающихся по своему расположению ССТФ в промоторах и энхансерах на уровень экспрессии генов является одним из важных направлений исследований; перекрывающиеся нуклеотидными последовательностями сайты связывания изучены недостаточно [15].

С использованием данных ChIP-seq для профилей связывания ТФ в геноме мыши ранее были исследованы взаимодействия транскрипционных факторов в плане одновременного связывания различных ТФ в геномных районах (так называемые множественные локусы регуляции транскрипции) [3; 12].

Одним из важных биомедицинских приложений является построение полногеномных карт регуляторов плюрипотентности NANOG, OCT4, SOX2, KLF4 в эмбриональных стволо-

вых клетках и связанных с ними кластеров сайтов других транскрипционных факторов [3]. Накоплены экспериментальные данные о трехмерной организации геномных участков (удаленные энхансеры, пространственные домены), что служит основой для более сложных моделей регуляторных районов [1; 4].

В настоящей работе представлены скрипты для анализа сайтов по данным ChIP-seq, расчета кластеров сайтов и их визуализации в форме тепловых карт, развивающие подходы, представленные в [12] на новых данных и выполненные в другой среде программирования.

Материалы и методы

В работе использовались пики ChIP-seq для клеточной линии эмбриональных стволовых клеток человека H1. Данные в виде BED-файлов загрузили из базы данных Cistrome¹ [16], были загружены координаты пиков 38 транскрипционных факторов.

Отметим, что существует огромное количество баз данных по сайтам связывания транскрипционных факторов, которые содержат в себе как данные по ChIP-seq (Expression Atlas, Roadmap epigenomics project, ENCODE), так и данные по непосредственным координатам ССТФ и мотивам связывания ТФ (TRANSFAC, JASPAR) [17; 18], также существуют базы данных, разработанные в России, – TRRD [19; 20], GTRD [21], HOCOMOCO [22].

Для анализа данных был разработан набор скриптов на языке R, в последующем они были собраны в пакет, названный ClanChIPeaks, который можно свободно скачать из репозитория GitHub². При разработке ClanChIPeaks и анализе данных также использовались сторонние пакеты из репозитория Bioconductor (GenomicRanges, AnnotationHub, ChIPeeker и т. д.) и CRAN (ggplot2, fastclust и т. д.). Общий алгоритм анализа пиков ChIP-seq при работе в пакете ClanChIPeaks и их кластеризация представлены на рис. 1.

На первом этапе необходимо загрузить экспериментальные данные в среду R при помощи функции `peaks.read()`, данная функция возвращает объект `GenomicRanges`. Отдельно этот объект и другие широко используемые классы объектов, используемые в Bioconductor для работы с биологическими данными, представлен в статье [23].

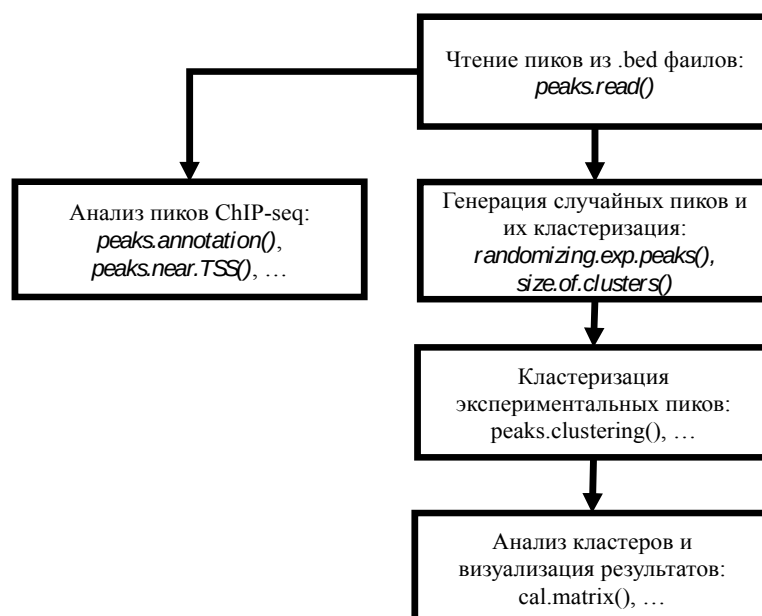


Рис. 1. Общий алгоритм анализа и кластеризации пиков ChIP-seq

¹ <http://cistrome.org/db/#/>.

² <https://github.com/anton-tsukanov/ClanChIPeaks>.

На втором этапе отдельно от анализа кластеров ClanChIPeaks позволяет проводить небольшой анализ пиков ChIP-seq. Так, например расчет плотности распределения пиков около начала сайтов транскрипции с использованием функции `peaks.near.TSS()`, а также аннотация пиков, т. е. в каком регионе находятся пики (экзон, интрон, 5'-нетранслируемая область, 3'-нетранслируемая область, межгенное пространство), для этого используется функция `peaks.annotation()`.

На третьем этапе осуществляется генерация случайных пиков ChIP-seq, которые обладают той же общей шириной пиков, и проводится их кластеризация. Данный этап необходим для выявления размера кластера, который не будет получаться по случайным причинам.

На четвертом этапе непосредственно проводится кластеризация пиков ChIP-seq полученных экспериментально, это осуществляется при помощи функции `peaks.clustering()`. Данная функция возвращает стандартный объект `list()`, где каждый элемент `list()` является объектом `data.frame()`, содержит пики из одного кластера и информацию о них (начало пика, конец пика, ширина и др.). При помощи функции `cal.matrix()` можно посчитать попарную встречаемость каждого транскрипционного фактора, результат будет представлен в виде матрицы.

На последнем этапе можно проводить анализ посчитанных данных при помощи стандартных методов, встроенных в R, а также визуализировать данные при помощи пакетов `ggplot2`, `corrplot` и др.

Результаты

Общий анализ пиков ChIP-seq

На основании геномных координат (начало и конец геномного участка) пиков профилей ChIP-seq для 38 транскрипционных факторов клеточной линии H1 была рассчитана ширина каждого пика для 9 ТФ (E2F6, CREB1, MYC, ZNF143, MXI1, SP4, YY1, NRF1, NANOG); построены графики распределения ширины пиков, представленные на рис. 2.

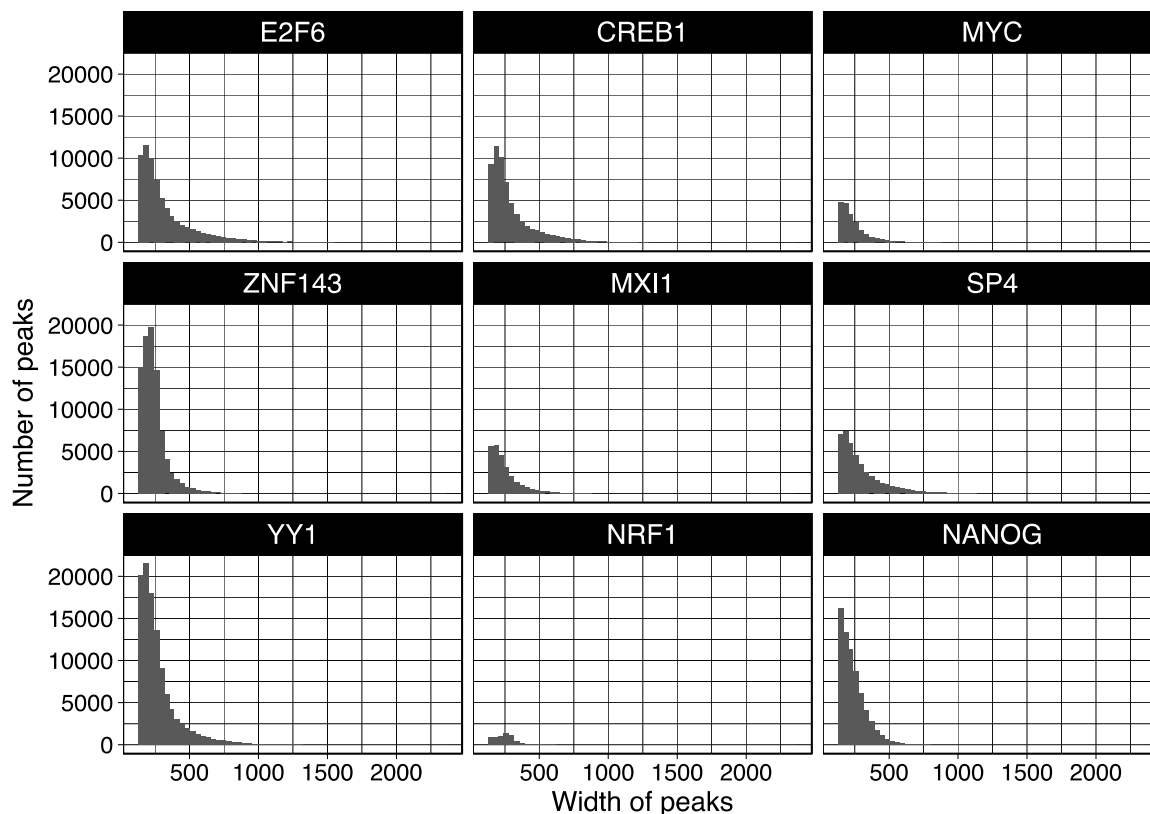


Рис. 2. Распределение размеров и ширины пиков ChIP-seq в геноме
Ось абсцисс – ширина пиков, ось ординат – количество пиков

Для представленных на рис. 2 транскрипционных факторов минимальное значение ширины пика составляет 147 п. н., медианные значения ширины пиков варьируют от 173 п. н. у MXI1 до 255 п. н. у NRF1, а максимальные значения ширины пиков – от 1074 п. н. у NANOG до 2345 п. н. у E2F6. Отметим, что сам сайт связывания ТФ имеет длину порядка 10 п. н., что делает необходимым искать точное положение сайтов связывания ТФ по их мотивам в нуклеотидных последовательностях пиков геномного профиля ChIP-seq размером в сотни нуклеотидов.

Также было посчитано распределение положения пиков ChIP-seq для 6 ТФ из общего набора в 38 ТФ клеточной линии H1 в разных участках генома относительно гена (интроны (Introns), экзоны (Exons), промоторы (Promoters), 5'-НТО (5UTR), 3'-НТО (3UTR), межгенное пространство (Intergenic)). Результаты представлены на рис. 3.

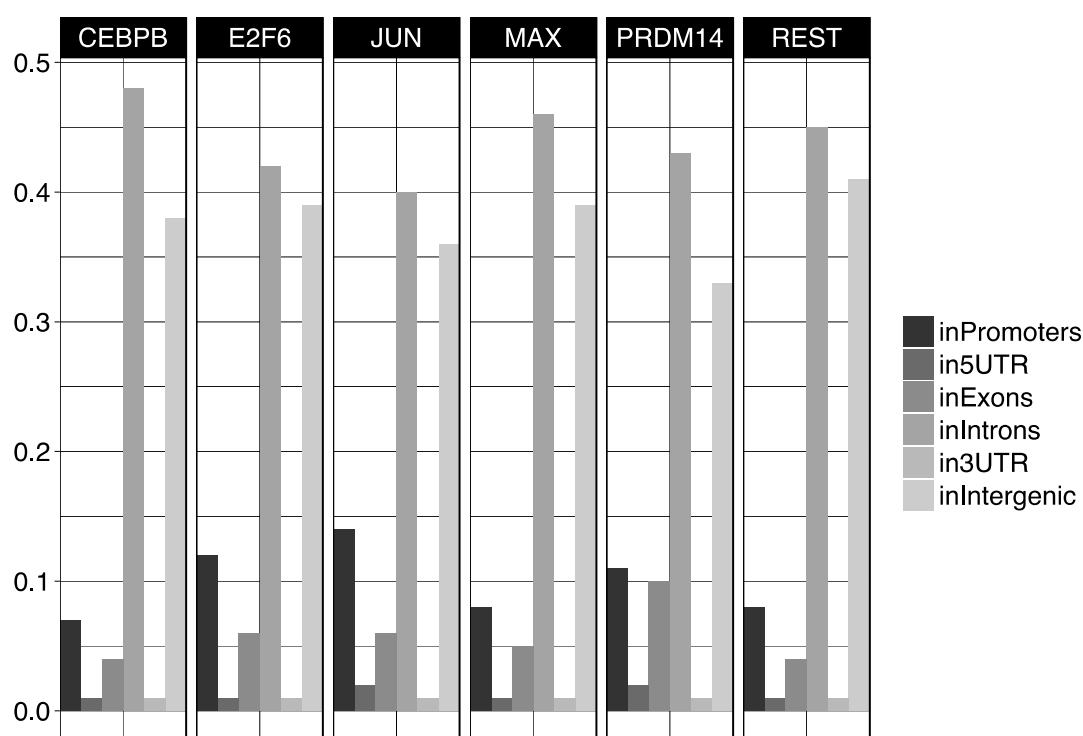


Рис. 3. Распределение пиков в геноме

Из рис. 3 видно, что большая часть сайтов находится в интронах и в межгенных районах. В то же время для ТФ JUN доля сайтов в промоторах выше, чем для ТФ CEBP и REST. Распределение сайтов в геноме дает лишь общую картину расположения сайтов относительно генов – наибольший интерес представляет определение положения сайта относительно старта транскрипции, что непосредственно влияет на транскрипцию данного гена.

Кластеризация сгенерированных пиков

Перед тем как изучать кластеризацию пиков ChIP-seq, полученных экспериментально, кластеризацию пиков изучали на сгенерированных данных, при этом суммарная ширина пиков – как экспериментальных, так и сгенерированных, совпадала. Случайные пики генерировали следующим образом. Зная общее число экспериментальных пиков на хромосоме, генерировали такое же количество псевдослучайных пиков (при помощи генератора псевдослучайных чисел) в диапазоне длины хромосомы. Далее сгенерированным пикам задавали значения ширины путем случайного выбора ширины из экспериментальных пиков.

Для кластеризации сгенерированных пиков использовали иерархический тип кластеризации и меру расстояния в Евклидовом пространстве. Кластеризацию проводили с использованием стороннего пакета *fastclust* из репозитория CRAN. Пики ChIP-seq кластеризовались друг с другом, если расстояние от ближайших центров пиков не превышало 25 п. н., таким образом размер кластера увеличивался до тех пор, пока расстояние между кластером и другим пиком не стало больше 25 п. н.

Сгенерированные данные ChIP-seq имели 1 485 464 пика, а точка отсечения составляла 25 п. н., для каждой новой кластеризации генерировались новые случайные пики ChIP-seq, всего было проведено 5 реплик. Результаты кластеризации сгенерированных пиков представлены в таблице.

Размер и количество кластеров,
полученных на сгенерированных пиках

Реплика	Размер кластера *			
	максимальный	количество кластеров	предмаксимальный	количество кластеров
1	4	5	3	267
2	4	6	3	263
3	4	6	3	263
4	4	4	3	294
5	4	5	3	283

* Размер кластера – количество пиков в кластере.

Из результатов, представленных в таблице, видно, что максимальный размер кластера достигает 4, а количество таких кластеров варьирует от 4 до 6. Таким образом, для экспериментальных данных ChIP-seq клеточной линии H1 (при условии, что для анализа мы используем 1 485 464 пика) мы приняли гипотезу о том, что минимальный размер кластера, который может образоваться не по случайным причинам, составляет 5 пиков ChIP-seq.

Кластеризация и изучение кластеров на экспериментальных данных ChIP-seq

Условия кластеризации пиков ChIP-seq, полученных экспериментально, были такие же, как и для сгенерированных пиков (иерархический тип кластеризации, мера расстояния – Евклидово пространство, максимальное расстояние между пиками в одном кластере 25 п. н.). В кластеризации использовались данные по 31 транскрипционному пику ChIP-seq для клеточной линии человека H1, общее количество пиков составило 1 485 464. Сопоставление размеров кластеров, полученных на экспериментальных данных ChIP-seq и сгенерированных, представлены на рис. 4. Из графика видно, что большая часть кластеров – это одиночные сайты и пары сайтов. Логарифм значения числа кластеров уменьшается с ростом количества пиков в самом кластере, при этом для экспериментальных кластеров, полученных из экспериментальных данных, вырисовывается S-образная кривая. Стоит отметить, что логарифм от значения количества кластеров плавно уменьшается только в диапазоне 1–30, а далее значение становится нестабильным – то уменьшается, то увеличивается (что связано с малым числом таких больших кластеров). Максимальный размер кластера достигает 37 пиков, а так как количество уникальных транскрипционных факторов всего 31, следовательно, в один кластер могли попадать пики одного и того же ТФ. Для случайно сгенерированных координат не наблюдалось кластеров размером больше 4. Таким образом, кластеры сайтов размером 4 и выше не случайны, что подтверждают сделанные ранее оценки [3; 12].

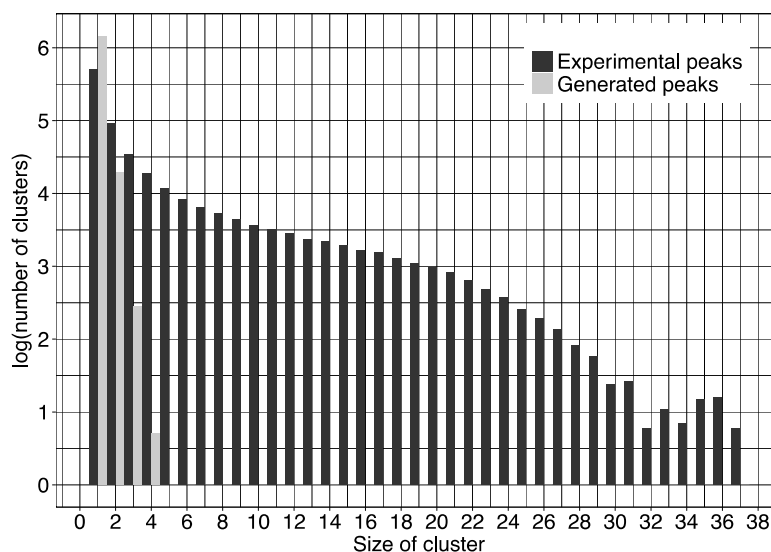


Рис. 4. Кластеры сайтов связывания ТФ (пиков ChIP-seq) по экспериментальным данным (черные столбцы) и компьютерная симуляция кластеров (генерация координат с помощью датчика случайных чисел)

Для дальнейшего анализа использовались только значимые кластеры, т. е. кластеры, размер которых составлял 5 пиков или больше. Для ТФ, входящих в такие кластеры, была посчитана матрица встречаемости, которая представляет собой квадратную симметричную матрицу чисел совместной встречаемости пары сайтов для каждого двух транскрипционных факторов из исследуемого набора. Из нее была посчитана матрица корреляций (по векторам целочисленных значений встречаемости пар сайтов). Все последующие подсчеты проводили на основе матрицы корреляций. Так, на основе матрицы корреляции был посчитан коэффициент несходства (различия), который считается как $(1 - \text{коэффициент корреляции})$, для каждой пары ТФ. Этот коэффициент использовался в качестве меры расстояния для построения дендрограммы (рис. 5), характеризующей взаимную встречаемость транскрипционных факторов.

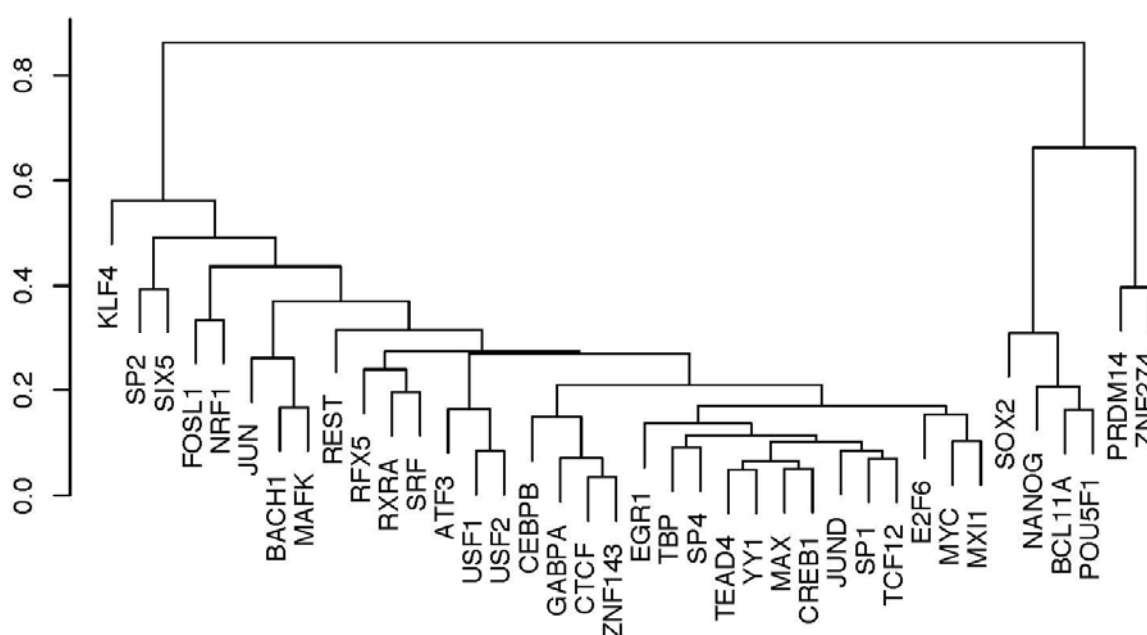


Рис. 5. Дендрограмма близости ТФ (пиков ChIP-seq) по корреляциям их взаимного расположения в кластерах

Из дендрограммы, представленной на рис. 5, можно сделать вывод, что некоторые ТФ предпочитают находиться в одном кластере. Такая, например, группа ТФ, как SOX2, NANOG, BCL11A и POU5F1, образует одну кладу, или PRDM14 и ZNF274, которые также образуют свою кладу.

Иным способом представления матрицы корреляции встречаемости ТФ является тепловая карта встречаемости ТФ (рис. 6).

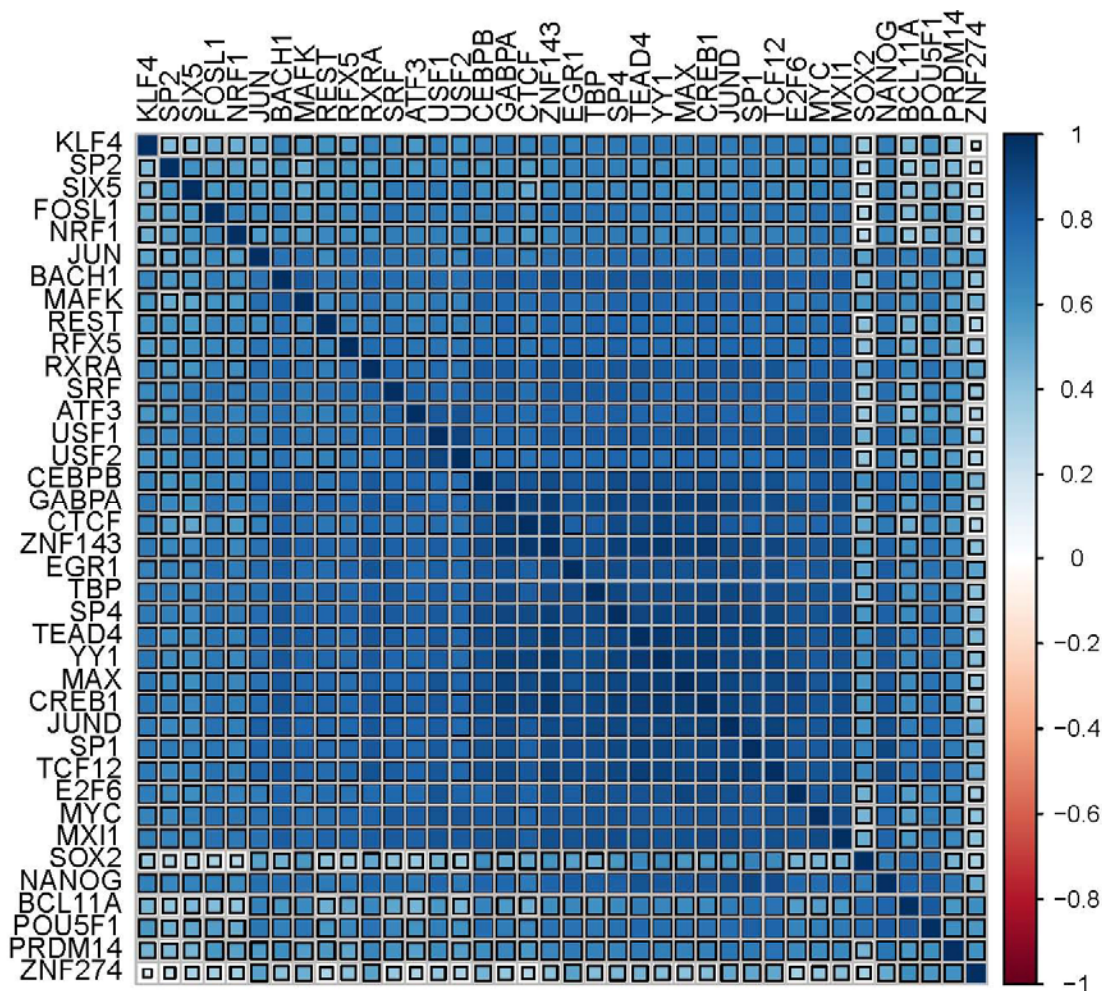


Рис. 6. Тепловая карта коэффициентов корреляции

Более темный синий цвет ячейки соответствует повышенной совместной встречаемости пары сайтов для транскрипционных факторов, представленных в соответствующих строке и столбце матрицы. Две основные ветви на дендрограмме (см. рис. 5), соответствующие факторам KLF4, SP2 и др. и факторам SOX2, NANOG и др., визуально выделяются как левый верхний и правый нижний более темные квадраты на тепловой карте (см. рис. 6). Тепловую карту построили с помощью пакета *corrplot*. И действительно, транскрипционные факторы SOX2, NANOG, POU5F1, PRDM14 из второго кластера являются факторами поддержания плюрипотентности, и совместная встречаемость их сайтов связывания обусловлена их общей функцией в клетке [3].

Заключение

Проведен вычислительный эксперимент по анализу кластеров сайтов связывания в геноме человека по базам данных ENCODE и Cistrome. Разработанные программы позволяют оценивать статистические параметры профилей ChIP-seq в геноме человека и модельных гено-

мах, строить распределения пиков по ширине и высоте (силе связывания с ДНК или уровнем значимости участка, представленным p -value), определять параметры отдельных групп сайтов. Программы позволяют рассчитывать распределение сайтов в геноме относительно старта транскрипции гена (для набора сайтов и разметки генов). Возможен статистический расчет положения сайтов в экзон-интронной структуре гена.

Программы позволяют рассчитывать колокализацию (совместную локализацию) сайтов связывания различных транскрипционных факторов, выполнять визуализацию, строить дендрограммы и тепловые карты совместной локализации сайтов.

Применение программ для анализа кластеров сайтов связывания в эмбриональных стволовых клетках человека по данным ChIP-seq из ресурса Cistrome позволило уточнить состав кластеров сайтов транскрипционных факторов, описать их функциональную роль. По частоте сигналов в исследованном наборе выделяются группы, относящиеся к NANOG, что подтверждает полученные ранее данные [3]. Среди 31 транскрипционного фактора такие группы факторов, как SOX2, NANOG, BCL11A, PRDM14 и POU5F1, имеют тенденцию встречаться совместно более часто.

Дальнейший анализ контекстных признаков в геномных последовательностях может опираться на участки низкой сложности текста (простые повторы и политракты), сайты связывания нуклеосом [24; 25]. Интеграция геномных данных позволяет решать качественно новые задачи, представляя описание полногеномной информации, такой как данные проектов ENCODE³, FactorBook⁴ в геноме человека [26]. Интересно отметить паттерны расположения участков простых нуклеотидных повторов (пониженной сложности текста) в районах однонуклеотидных полиморфизмов в геноме человека [27]. Участки простых повторов в геноме труднее картировать в геноме по коротким последовательностям прочтений ДНК при секвенировании [28]. Анализ этих сигналов вокруг кластеров сайтов связывания транскрипционных факторов позволит построить модель организации таких геномных районов [13; 14], предсказать их функцию по составу сайтов и контекстным характеристикам. В целом данное исследование кластеров сайтов связывания транскрипционных факторов в геноме развивает применение анализа регуляции экспрессии генов в эмбриональных стволовых клетках [29].

Список литературы

1. Игнатьева Е. В., Подколотная О. А., Орлов Ю. Л., Васильев Г. В., Колчанов Н. А. Регуляторная геномика – экспериментально-компьютерные подходы // Генетика. 2015. Т. 51 (4). С. 409–429.
2. Levine M., Cattoglio C., Tjian R. Looping back to leap forward: transcription enters a new era // Cell. 2014. No. 157. P. 13–25.
3. Chen X., Xu H., Yuan P. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells // Cell. 2008. Vol. 133. No. 6. P. 1106–1117.
4. Кулакова Е. В., Спицина А. М., Орлова Н. Г., Дергилев А. И., Свичкарев А. В., Сафронова Н. С., Черных И. Г., Орлов Ю. Л. Программы анализа геномных данных секвенирования, полученных на основе технологий ChIP-seq, ChIA-PET и Hi-C // Программные системы: теория и приложения. 2015. Т. 6. № 2 (25). С. 129–148.
5. Fitzgerald K. A. et al. The role of transcription factors in prostate cancer and potential for future RNA interference therapy // Nucleic Acids Research. 2015. Vol. 43. No. 14. P. 6874–6888.
6. Zhu H., Wang G., Qian J. Transcription factors as readers and effectors of DNA methylation // Nature. 2016. Vol. 17. P. 551–565.
7. Kelly T. K., Liu Y., Lay F. D., Liang G., Berman B. P., Jones P. A. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules // Genome Research. 2012. No. 22. P. 2497–2506.
8. Hu Z., Tee W. Enhancers and chromatin structures: regulatory hubs in gene expression and diseases // Bioscience Reports. 2017. No. 37. P. 1–14.

³ <https://genome.ucsc.edu/ENCODE/>.

⁴ <http://www.factorbook.org>.

9. *Guillaume A., Stefan M.* The three-dimensional genome: regulating gene expression during pluripotency and development // *Development*. 2017. Vol. 144. P. 3646–3658.
10. Орлов Ю. Л., Тьерри О., Богомолов А. Г., Цуканов А. В., Кулакова Е. В., Галиева Э. Р., Брагин А. О., Ли Г. Компьютерные методы анализа хромосомных контактов в ядре клетки по данным технологий секвенирования // *Биомедицинская химия*. 2017. № 63 (5). С. 418–422.
11. Орлов Ю. Л., Брагин А. О., Медведева И. В., Гунбин И. В., Деменков П. С., Вишневский О. В., Левицкий В. Г., Ощепков В. Г., Подколотный Н. Л., Афонников Д. А., Гроссе И., Колчанов Н. А. ICGenomics: программный комплекс анализа символьных последовательностей геномики // *Вавиловский журнал генетики и селекции*. 2012. Т. 16, № 4/1. С. 732–741.
12. Дергулев А. И., Спицина А. М., Чадаева И. В., Свичкарев А. В., Науменко Ф. М., Кулакова Е. В., Витяев Е. Е., Чен М., Орлов Ю. Л. Компьютерный анализ совместной локализации сайтов связывания транскрипционных факторов по данным ChIP-seq // *Вавиловский журнал генетики и селекции*. 2016. Т. 20 (6). С. 770–778. DOI 10.18699/VJ16.194.
13. *Vityaev E. E., Orlov Yu. L., Vishnevsky O. V., Pozdnyakov M. A., Kolchanov N. A.* Computer system «Gene Discovery» for promoter structure analysis // *In Silico Biology*. 2002. Vol. 2. No. 3. P. 233–247.
14. Витяев Е. Е., Орлов Ю. Л., Вишневский О. В., Беленок А. С., Колчанов Н. А. Компьютерная система «Gene Discovery» для поиска закономерностей организации регуляторных последовательностей эукариот // *Молекулярная биология*. 2001. Т. 35, В 6. С. 952–960.
15. Васькин Ю. Ю., Хомичева И. В., Игнатьева Е. В., Витяев Е. Е. Анализ последовательностей регуляторных районов генов реляционной системой ExpertDiscovery, встроенной в пакет UGENE // *Вестн. НГУ. Серия: Информационные технологии*. 2012. Т. 10. № 1. С. 73–86.
16. *Mei S., Qin Q., Wu Q., Sun H., Zheng R., Zang C., Zhu M., Wu J., Shi X., Taing L., Liu T., Brown M., Meyer C. A., Liu X. S.* Cistrome data browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse // *Nucleic Acids Res*. 2017. Vol. 45. No. 4. P. 658–662.
17. *Knuppel R., Dietze P., Lehnberg W., Frech K., Wingender E.* TRANSFAC1 retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins // *J. Comput. Biol*. 1994. Vol. 1. P. 191–198.
18. *Mathelier A. et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles // *Nucleic Acids Res*. 2016. Vol. 44. P. 110–115.
19. Кель А. Э., Колчанов Н. А., Кель О. В., Ромащенко А. Г., Ананько Е. А., Игнатьева Е. В., Меркулова Т. И., Подколотная О. А., Степаненко И. Л., Кочетов А. В., Колпаков Ф. А., Подколотный Н. Л., Наумочкин А. А. TRRD: база данных транскрипционных регуляторных районов генов эукариот // *Молекулярная биология*. 1997. Т. 31, № 4. С. 636–672.
20. *Kolchanov N. A., Ignatieva E. V., Ananko E. A., Podkolodnaya O. A., Stepanenko I. L., Merkulova T. I., Pozdnyakov M. A., Podkolodny N. L., Naumochkin A. N., Romashchenko A. G.* Transcription Regulatory Regions Database (TRRD): its status in 2002 // *Nucleic Acids Res*. 2002. Vol. 30 (1). P. 312–7.
21. *Yevshin I., Sharipov R., Valeev T., Kel A., Kolpakov F.* GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments // *Nucleic Acids Res*. 2017. Vol. 45 (D1). P. D61–D67.
22. *Kulakovskiy I. V., Vorontsov I. E., Yevshin I. S., Sharipov R. N., Fedorova A. D., Rumynskiy E. I., Medvedeva Y. A., Magana-Mora A., Bajic V. B., Papatsenko D. A., Kolpakov F. A., Makeev V. J.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis // *Nucleic Acids Res*. 2018. Vol. 46 (D1). P. D252–D259. DOI: 10.1093/nar/gkx1106.
23. *Lawrence M. et al.* Software for Computing and Annotating Genomic Ranges // *PLOS Computational Biology*. 2013. Vol. 8. P. 1–10.
24. *Orlov Yu. L., Potapov V. N.* Complexity: an internet resource for analysis of DNA sequence complexity // *Nucleic Acids Res*. 2004. Vol. 32. P. W628–W633.
25. Орлов Ю. Л., Левицкий В. Г., Смирнова О. Г., Подколотная О. А., Хлебодарова Т. М., Колчанов Н. А. Статистический анализ последовательностей ДНК, содержащих сайты формирования нуклеосом // *Биофизика*. 2006. Т. 51. С. 608–614.

26. Спицина А. М., Орлов Ю. Л., Подколотная Н. Н., Свичкарев А. В., Дергилев А. И., Чен М., Кучин Н. В., Черных И. Г., Глинский Б. М. Суперкомпьютерный анализ геномных и транскриптомных данных, полученных с помощью технологий высокопроизводительного секвенирования ДНК // Программные системы: теория и приложения. 2015. Т. 6, № 1 (23). С. 157–174.

27. Сафронова Н. С., Пономаренко М. П., Абнизова И. И., Орлова Г. В., Чадаева И. В., Орлов Ю. Л. Фланкирующие повторы мономеров определяют пониженную контекстную сложность сайтов однонуклеотидных полиморфизмов в геноме человека // Вавиловский журнал генетики и селекции. 2015. Т. 19 (6). С. 668–674.

28. Naumenko F. M., Abnizova I. I., Beka N., Genaev M. A., Orlov Yu. L. Novel read density distribution score shows possible aligner artefacts, when mapping a single chromosome // BMC Genomics. 2018. Vol. 19 (Suppl. 3). P. 92. DOI: 10.1186/s12864-018-4475-6/

29. Дергилев А. И., Цуканов А. В., Орлов Ю. Л. Компьютерный анализ кластеров сайтов связывания транскрипционных факторов в эмбриональных стволовых клетках // Гены и клетки. 2017. Т. 12 (3). С. 184–185.

Материал поступил в редколлегию 02.07.2018

A. V. Tsukanov^{1,2}, N. G. Orlova³, A. I. Dergilev², Yu. L. Orlov^{1,2}

¹ *Institute of Cytology and Genetics SB RAS
10 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation*

² *Novosibirsk State University
1 Pirogov Str., Novosibirsk, 630090, Russian Federation*

³ *Novosibirsk State University of Architecture and Civil Engineering (Sibstrin)
113 Leningradskaya Str., Novosibirsk, 630008, Russian Federation*

ya.cukanton@yandex.ru, orlovanina2@mail.ru, arturd1993@yandex.ru, orlov@bionet.nsc.ru

PROGRAMS FOR STATISTICAL ANALYSIS, CLUSTERIZATION AND VISUALIZATION OF GENOME DISTRIBUTION OF TRANSCRIPTION FACTOR BINDING SITES

The analysis of gene transcription regulation based on the data of modern technologies of high-performance sequencing is an actual task of bioinformatics. It requires the development of new computer tools including supercomputer applications. We consider the problems of processing of genome ChIP-seq profiles for detections of transcription factors binding site in a genome, determining the peaks of such profiles and search the binding sites in the nucleotide sequences of the peaks. The computer programs have been developed to analyze the location of the binding sites in the genome relative to gene regions, to calculate clusters of such sites and visualize their positions in the genome. Clusters of binding sites of transcription factors in the human genome have been calculated using the Cistrome database. We have calculated matrices of the joint occurrence of pairs of binding sites of different transcription factors in the genome for various types of tissues and cells. A computational experiment on the computer generation of random clusters in the genome was carried out, as well as an assessment of the occurrence of large clusters for experimentally obtained binding sites of transcription factors in the human genome. The patterns of occurrence of binding sites of pluripotency factors in embryonic stem cells were described. The developed software is available on request to the authors.

Keywords: genomics, sequencing, binding sites of transcription factors, promoters, big data, statistics, visualization.

References

1. Ignatieva E. V., Podkolodnaya O. A., Orlov Yu. L., Vasiliev G. V., Kolchanov N. A. Regulatory genomics: combined experimental and computational approaches. *Russian Journal of Genetics*, 2015, vol. 51 (4), p. 409–429. (in Russ.)
2. Levine M., Cattoglio C., Tjian R. Looping back to leap forward: transcription enters a new era. *Cell*, 2014, no. 157, p. 13–25.
3. Chen X., Xu H., Yuan P. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 2008, vol. 133, no. 6, p. 1106–1117.
4. Kulakova E. V., Spitsina A. M., Orlova N. G., Dergilev A. I., Svichkarev A. V., Safronova N. S., Chernykh N. S., Orlov Yu. L. Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing. *Program systems: theory and applications*, 2015, vol. 6, no. 2 (25), p. 129–148. (in Russ.)
5. Fitzgerald K. A. et al. The role of transcription factors in prostate cancer and potential for future RNA interference therapy. *Nucleic Acids Research*, 2015, vol. 43, no. 14, p. 6874–6888.
6. Zhu H., Wang G., Qian J. Transcription factors as readers and effectors of DNA methylation. *Nature*, 2016, vol. 17, p. 551–565.
7. Kelly T. K., Liu Y., Lay F. D., Liang G., Berman B. P., Jones P. A. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Research*, 2012, no. 22, p. 2497–2506.
8. Hu Z., Tee W. Enhancers and chromatin structures: regulatory hubs in gene expression and diseases. *Bioscience Reports*, 2017, no. 37, p. 1–14.
9. Guillaume A., Stefan M. The three-dimensional genome: regulating gene expression during pluripotency and development. *Development*, 2017, vol. 144, p. 3646–3658.
10. Orlov Yu. L., Thierry O., Bogomolov A. G., Tsukanov A. V., Kulakova E. V., Galieva E. R., Bragin A. O., Li G. Computer methods of analysis of chromosome contacts in the cell nucleus based on sequencing technology data. *Biomeditsinskaya Khimiya*, 2017, no. 63 (5), p. 418–422. (in Russ.)
11. 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. ICGenomics: a program complex for analysis of symbol sequences in genomics. *Vavilovskii Zhurnal Genetiki i Selektzii = Vavilov Journal of Genetics and Breeding*, 2012, vol. 16, no. 4/1, p. 732–741. (in Russ.)
12. Dergilev A. I., Spitsina A. M., Chadaeva I. V., Svichkarev A. V., Naumenko F. M., Kulakova E. V., Vityaev E. E., Chen M., Orlov Yu. L. Computer analysis of colocalization of the TFs' binding sites in the genome according to the ChIP-seq data. *Russian Journal of Genetics: Applied Research*, 2016, vol. 20 (6), p. 770–778. DOI 10.18699/VJ16.194. (in Russ.)
13. Vityaev E. E., Orlov Yu. L., Vishnevsky O. V., Pozdnyakov M. A., Kolchanov N. A. Computer system «Gene Discovery» for promoter structure analysis. *In Silico Biology*, 2002, vol. 2, no. 3, p. 233–247.
14. 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. *Molecular Biology*, 2001, vol. 35, B 6, p. 952–960. (in Russ.)
15. Vaskin Yu. Yu., Khomicheva I. V., Ignatyeva E. V., Vityayev E. E. Analysis of regulatory regions of genes by Expert Discovery relation system, integrated into UGENE toolkit. *Vestnik NSU. Series: Information Technologies*, 2012, vol. 10, no. 1, p. 73–86. (in Russ.)
16. Mei S., Qin Q., Wu Q., Sun H., Zheng R., Zang C., Zhu M., Wu J., Shi X., Taing L., Liu T., Brown M., Meyer C. A., Liu X. S. Cistrome data browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, 2017, vol. 45, no. 4, p. 658–662.
17. Knuppel R., Dietze P., Lehnberg W., Frech K., Wingender E. TRANSFAC1 retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. *J. Comput. Biol.*, 1994, vol. 1, p. 191–198.
18. Mathelier A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 2016, vol. 44, p. 110–115.

19. Kel A. E., Kolchanov N. A., Kel O. V., Romashchenko A. G., Ananko E. A., Ignatieva E. V., Merkulova T. I., Podkolodnaya O. A., Stepanenko I. L., Kochetov A. V., Kolpakov F. A., Podkolodnyi N. L., Naumochkin A. A. TRRD: a database of transcription regulatory regions in eukaryotic genes. *Mol. Biol.*, 1997, vol. 31, no. 4, p. 636–672. (in Russ.)
20. Kolchanov N. A., Ignatieva E. V., Ananko E. A., Podkolodnaya O. A., Stepanenko I. L., Merkulova T. I., Pozdnyakov M. A., Podkolodny N. L., Naumochkin A. N., Romashchenko A. G. Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res.*, 2002, vol. 30 (1), p. 312–7.
21. Yevshin I., Sharipov R., Valeev T., Kel A., Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.*, 2017, vol. 45 (D1), p. D61–D67.
22. Kulakovskiy I. V., Vorontsov I. E., Yevshin I. S., Sharipov R. N., Fedorova A. D., Rumynskiy E. I., Medvedeva Y. A., Magana-Mora A., Bajic V. B., Papatsenko D. A., Kolpakov F. A., Makeev V. J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, 2018, vol. 46 (D1), p. D252–D259. DOI: 10.1093/nar/gkx1106.
23. Lawrence M. et al. Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*, 2013, vol. 8, p. 1–10.
24. Orlov Yu. L., Potapov V. N. Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic Acids Res.*, 2004, vol. 32, p. W628–W633.
25. Orlov Yu. L., Levitskii V. G., Smirnova O. G., Podkolodnaya O. A., Khlebodarova T. M., Kolchanov N. A. Statistical analysis of DNA sequences containing nucleosome positioning sites. *Biophysics*, 2006, vol. 51, p. 608–614. (in Russ.)
26. Spitsina A. M., Orlov Yu. L., Podkolodnaya N. N., Svicharev A. V., Dergilev A. I., Chen M., Kuchin N. V., Chernykh I. G., Glinsky B. M. Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing. *Program systems: theory and applications*, 2015, vol. 6, no. 1 (23), p. 157–174. (in Russ.)
27. Safronova N. S., Ponomarenko M. P., Abnizova I. I., Orlova G. V., Chadaeva I. V., Orlov Yu. L. Flanking monomer repeats determine decreased context complexity of single nucleotide polymorphism sites in the human genome. *Russian Journal of Genetics: Applied Research*, 2015, vol. 19 (6), p. 668–674. (in Russ.)
28. Naumenko F. M., Abnizova I. I., Beka N., Genaev M. A., Orlov Yu. L. Novel read density distribution score shows possible aligner artefacts, when mapping a single chromosome. *BMC Genomics*, 2018, vol. 19 (suppl. 3), p. 92. DOI: 10.1186/s12864-018-4475-6/
29. Dergilev A. I., Tsukanov A. V., Orlov Yu. L. Computer analysis of clusters of transcription factor binding sites in embryonic stem cells. *Genes and Cells*, 2017, vol. 12 (3), p. 184–185. (in Russ.)

For citation:

Tsukanov A. V., Orlova N. G., Dergilev A. I., Orlov Yu. L. Programs for Statistical Analysis, Clusterization and Visualization of Genome Distribution of Transcription Factor Binding Sites. *Vestnik NSU. Series: Information Technologies*, 2018, vol. 16, no. 3, p. 51–63. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-51-63