Письма в ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ
Letters to Vavilov Journal of Genetics and Breeding
2025 • 11 • 4 • 164-173

pismavavilov.ru
doi 10.18699/letvjgb-2025-11-23

Оригинальное исследование

MultiDeCA: конвейер для поиска и анализа мотивов сайтов связывания транскрипционных факторов на основе ChIP-seq данных

А.В. Цуканов ✉, В.Г. Левицкий

Аннотация. Регуляция транскрипции генов эукариот контролируется транскрипционными факторами (ТФ), распознающими специфические последовательности ДНК, сайты связывания ТФ (ССТФ) в регуляторных районах генов. Поиск ССТФ необходим для выяснения механизмов регуляции транскрипции генов. Технология массового секвенирования ChIP-seq позволяет картировать районы ССТФ в масштабах генома, что радикально расширяет возможности изучения ДНК-белковых взаимодействий. Анализ данных ChIP-seq включает следующие этапы обработки: картирование прочтений; поиск пиков – геномных локусов, где ТФ был связан с ДНК; *de novo* поиск мотивов как обогащенных паттернов нуклеотидного контекста для комплексов ТФ-ДНК в этих локусах; анализ полученных мотивов ССТФ. Последние два этапа наиболее важны, поэтому мы предлагаем конвейер MultiDeCA (Multi-model De novo discovery of motifs and their Combinatorial Analysis), включающий следующие инструменты. BSMotif кластеризует ТФ по сходству мотивов ССТФ на основе классификации ТФ по структуре ДНК-связывающего домена, AntiNoise готовит негативные выборки последовательностей ДНК для *de novo* поиска, MultiDeNA единообразно применяет модели мотивов PWM, BaMM, InMoDe и SiteGA. MCOT/MetArea находят совместную/взаимоисключающую встречаемость мотивов в парах. Инструменты BSMotif/AntiNoise/MultiDeNA позволяют упростить указание вовлеченных ТФ/повысить точность поиска мотивов ССТФ/отразить разнообразие выявляемых ССТФ. Инструменты MCOT/MetArea обеспечивают дополнительный анализ пар мотивов, выявляя потенциальных ТФ партнеров/непрямое связывание для целевого ТФ. Пример анализа ChIP-seq данных для ТФ JunD подтверждает эффективность конвейера.

Ключевые слова: сайты связывания транскрипционных факторов; ССТФ; ChIP-seq; мотивы; *de novo* поиск; биоинформатический конвейер

Для цитирования: Цуканов А.В., Левицкий В.Г. MiltiDeCA: конвейер для поиска и анализа мотивов сайтов связывания транскрипционных факторов на основе ChIP-seq данных. *Письма в Вавиловский журнал генетики и селекции*. 2025;11(4):164-173. doi 10.18699/letvjgb-2025-11-23

Финансирование. Работа поддержана бюджетным проектом FWNR-2022-0020.

Original article

MultiDeCA: pipeline for discovery and analysis of transcription factor binding site motifs based on ChIP-seq data

A.V. Tsukanov ✉, V.G. Levitsky

Abstract. The regulation of eukaryotic gene transcription is controlled by transcription factors (TFs) recognizing specific DNA sequences, TF binding sites (TFBSs) in the regulatory regions of genes. The TFBS search is essential to clarify the mechanisms of gene transcription regulation. The high-throughput sequencing technology ChIP-seq allows to map TFBSs genome-wide that radically expands the abilities for DNA-protein interaction study. The ChIP-seq data analysis includes the following processing steps: read mapping; peak calling implying detection of genomic loci where a TF bound DNA; *de novo* motif search identifying the enriched nucleotide context patterns for TF-DNA complexes in these loci; and the analysis of obtained TFBS motifs. The last two steps are the most important, thereby we propose the pipeline MultiDeCA (Multi-model De novo discovery of motifs and their Combinatorial Analysis) including the following tools. BSMotif tool clusters TFs by TFBS motif similarity based on TF classifications by DNA-binding domain structure. AntiNoise tool prepares the negative DNA sequences for *de novo* motif discovery. MultiDeNA tool uniformly applies the motif models PWM, BaMM, InMoDe, and SiteGA. MCOT/MetArea tools detect the motif co-occurrence/mutual exclusive occurrence in motif pairs. BSMotif/AntiNoise/MultiDeNA tools allow simplifying the detection of involved TFs/to increase the TFBS motif search accuracy/reflect the variety of detected TFBS motifs. MCOT/MetArea tools provide additional motif analysis to reveal the potential partner TFs/indirect binding of the target TF. The pipeline was successfully applied for the analysis of the target TF JunD ChIP-seq data.

Key words: transcription factors binding sites; TFBSs; ChIP-seq; motifs; *de novo* search; bioinformatics pipeline

For citation: Tsukanov A.V., Levitsky V.G. MultiDeCA: pipeline for discovery and analysis of transcription factor binding site motifs based on ChIP-seq data. *Pisma v Vavilovskii Zhurnal Genetiki i Selektsii = Lett Vavilov J Genet Breed.* 2025;11(4):164-173. doi 10.18699/letvjgb-2025-11-23 (in Russian)

Funding. This work was supported by State Budgeted Project FWNR-2022-0020.

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

tsukanov@bionet.nsc.ru

© Цуканов А.В., Левицкий В.Г., 2025

Контент доступен под лицензией Creative Commons Attribution 4.0

Введение

Регуляция транскрипции генов в эукариотических клетках осуществляется транскрипционными факторами (ТФ), которые связываются со специфическими последовательностями ДНК, сайтами связывания ТФ (ССТФ), и контролируют экспрессию генов (Lambert et al., 2018). ССТФ – участки ДНК от 6 до 20 пар нуклеотидов (п. н.). ТФ способны связываться со многими сходными сайтами связывания. Схожие сайты связывания одного ТФ представляют его мотив. Транскрипционный фактор может иметь два и более мотива. Способность ТФ взаимодействовать с ДНК определяется его ДНК-связывающим доменом (ДСД); ТФ разделяют по топологии ДСД на классы, выравнивания ДСД ТФ делят классы на семейства (база данных TFClass) (Wingender et al., 2013, 2015, 2018). Понимание механизмов распознавания ТФ их сайтов связывания критически важно для понимания механизмов регуляции транскрипции генов.

Появление и массовое применение технологии иммунопреципитации хроматина с последующим массовым секвенированием (ChIP-seq) произвели 10–15 лет назад революцию в изучении ДНК-белковых взаимодействий, предоставив возможность картирования локусов сайтов связывания ТФ в геномном масштабе (Farnham, 2009; Furey, 2012). ChIP-seq эксперименты дают тысячи геномных локусов (пиков) – районов связывания исследуемого ТФ, что требует применения специализированных биоинформатических подходов для их анализа.

Одной из ключевых задач анализа ChIP-seq данных является *de novo* поиск мотивов сайтов связывания ТФ – процесс выявления обогащенных мотивов ССТФ, изучаемых в экспериментах ChIP-seq, и других (партнерских) ТФ, вовлеченных в связывание целевых ТФ (Zambelli et al., 2013; Lihu, Holban, 2015). Традиционный способ представления мотива ССТФ – модель позиционной весовой матрицы (position weight matrix, PWM) (Stormo, 2000). Эта модель предполагает, что вклад каждой позиции сайта в общую оценку аффинности является независимым. Однако альтернативные модели мотива SiteGA (Levitsky et al., 2007; Tsukanov et al., 2022), BaMM (Siebert, Söding, 2016; Ge et al., 2021), InMoDe (Eggeling et al., 2017) могут не только достигать более высокой точности в предсказании ССТФ, но и полнее представлять структурное разнообразие мотивов ССТФ (Tsukanov et al., 2021, 2022).

Существующие инструменты анализа ChIP-seq данных часто ограничены в своих возможностях интеграции различных подходов к поиску мотивов и не обеспечивают комплексного анализа от исходных данных до биологической интерпретации результатов. Цель данной работы – представление конвейера MultiDeCA (Multi-model De novo discovery of motifs and their Combinatorial Analysis), объединяющего несколько специализированных инструментов для комплексной обработки ChIP-seq данных с поиском и анализом мотивов ССТФ.

Для *de novo* поиска мотивов критически важен выбор негативных последовательностей. Именно от него зависит оценка обогащения мотива, которая рассчитывается путем сравнения частоты его встречаемости в позитивной (пики ChIP-seq) и негативной выборках. Однако часто применяемые конвейеры используют слишком простые подходы для определения негативных выборок (Raditsa et al., 2024).

Стандартный подход корректного определения обогащения мотивов ССТФ, выявленных по данным ChIP-seq, представляет предварительная процедура перекрестной проверки для оценки точности распознавания мотивов (Simcha et al., 2012). С помощью этой процедуры можно выявлять набор надежных обогащенных мотивов при массовом анализе данных ChIP-seq (Tsukanov et al., 2022), процедура интегрирована в инструмент MultiDeNA (Tsukanov et al., 2021, 2022). Более того, этот инструмент позволяет сочетать несколько методологически разных моделей мотива, что позволяет существенно расширить количество пиков ChIP-seq с предсказанными ССТФ. После того как мотивы выявлены, необходимо сравнить их с мотивами сайтов связывания известных ТФ из баз данных, таких как JASPAR (Rauluseviciute et al., 2024) и HOCOMOCO (Vorontsov et al., 2024), это позволит оценить, выявлен ли мотив целевого ТФ или мотив другого ТФ. Обычно этот этап осложняется тем, что количество известных существенно различных мотивов ССТФ (от 100 до 300, Castro-Mondragon et al., 2017; Pratt et al., 2022; Rauluseviciute et al., 2024) существенно ниже, чем количество ТФ (около 1600 ТФ в геноме человека, Lambert et al., 2018), поэтому мы предварительно кластеризуем мотивы сайтов связывания известных ТФ по сходству с помощью инструмента BSMotif (Levitsky et al., 2025b), с учетом иерархической классификации ТФ по структуре ДСД (Wingender et al., 2013, 2015, 2018).

Возможно связывание целевого ТФ совместно с некоторым партнерским ТФ, тогда их ССТФ будут в пике рядом, или непрямое связывание целевого ТФ (целевой и партнерский ТФ имеют белок-белковое взаимодействие, но напрямую с ДНК связан только партнерский ТФ). Так как инструмент MultiDeNA выявляет мотивы сайтов связывания целевого ТФ, то для анализа партнерских ТФ мы применяем инструменты MCOT (Levitsky et al., 2019, 2020, 2025a) и MetArea (Levitsky et al., 2024). Совместная встречаемость двух обогащенных мотивов, выявляемая инструментом MCOT, отражает их присутствие в одних пиках; предполагается участие двух ТФ в одном мультибелковом комплексе, что может отражать синергию или антагонизм ТФ (Kel et al., 1995; Levitsky et al., 2019, 2020, 2025a). Взаимоисключающая встречаемость, выявляемая инструментом MetArea, разносит два обогащенных мотива в разные пики; предполагается, что два мотива относятся: а) к одному ТФ, способному к связыванию с различными мотивами, или б) к разным ТФ, заменяющим друг друга в составе мультибелкового комплекса, что может отражать антагонизм ТФ (Levitsky et al., 2024).

Методы

Предлагаемый конвейер MultiDeCA (рис. 1) представляет собой комплексное решение для анализа ChIP-seq данных, обеспечивающее выявление наиболее полного набора мотивов ССТФ. MultiDeCA объединяет несколько специализированных биоинформатических инструментов, каждый из которых решает специфические задачи на различных этапах анализа данных.

В отличие от существующих решений, использующих единственную модель мотива, предложенный конвейер обеспечивает систематический подход к выявлению всего

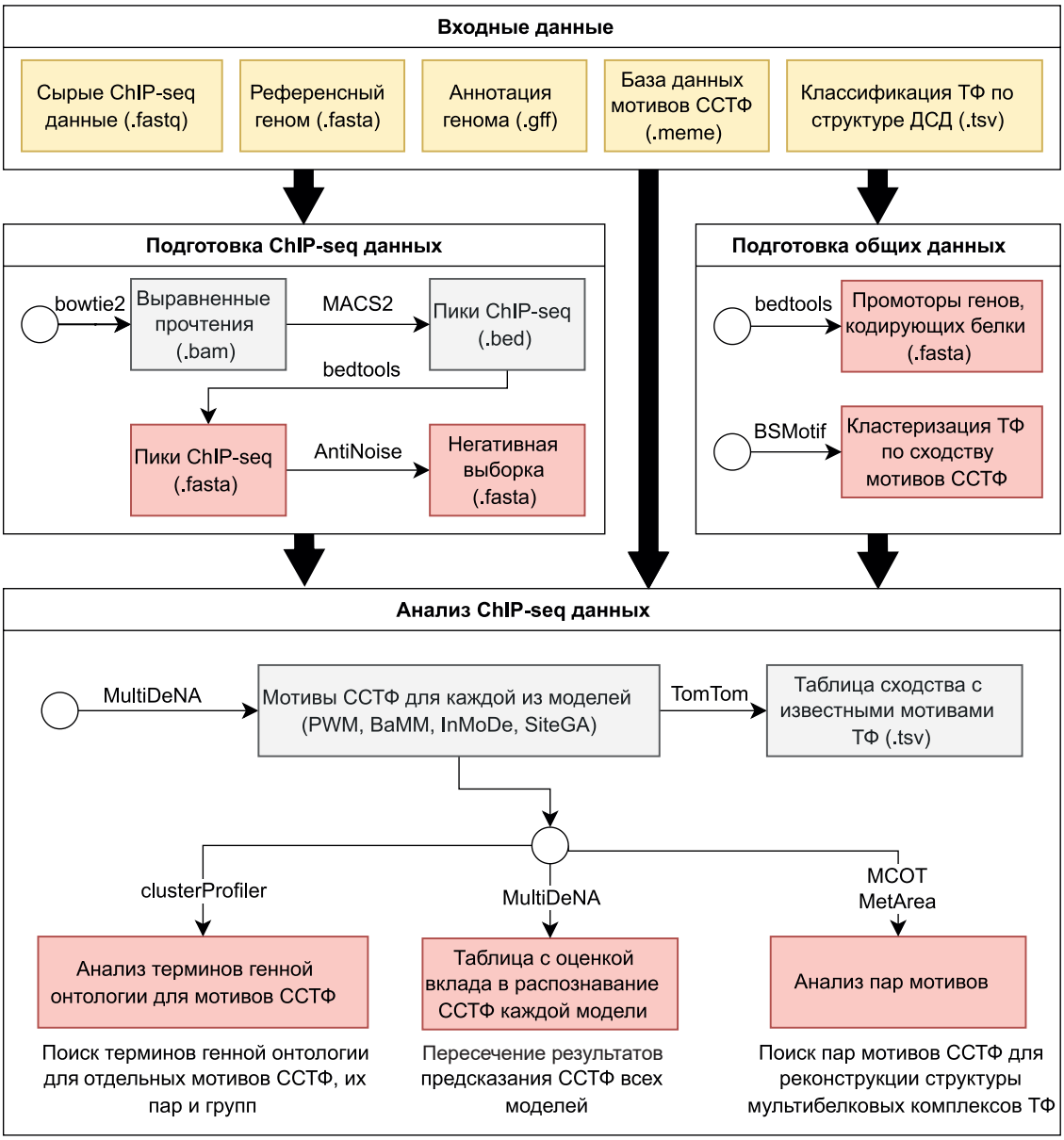

Рис. 1. Схема конвейера MultiDeCA. Цветом обозначены виды данных: желтый – входные данные от пользователя; серый – промежуточные данные, формируемые в процессе работы конвейера; красный – выходные данные или результаты отдельных блоков. Толстыми/тонкими стрелками показаны переход данных между разными блоками/обработка некоторым инструментом (анализ или преобразование).
Fig. 1. MultiDeCA scheme. The color indicates the data types: yellow means input data from user, gray denotes intermediate data forming in the conveyor, red implies output data or results of separate blocks. Thick/thin arrows show the transition of data between different blocks/data processing by certain tool (analysis or transformation).

спектра структурного разнообразия сайтов связывания. Кроме того, он позволяет аннотировать обнаруженные мотивы ССТФ, предсказывать их потенциальные биологические функции и оценивать взаимодействия между вовлеченными ТФ. Для этого конвейер включает анализ совместной/взаимоисключающей встречаемости мотивов, что может раскрыть механизмы функционирования ТФ в составе мультибелковых комплексов (Levitsky et al., 2019, 2020, 2024, 2025a).

Входные данные. Для анализа данных ChIP-seq конвейер принимает ряд входных данных, заданных файлами в следующих форматах: сырые данные ChIP-seq в формате FASTQ; референсный геном в формате FASTA; аннотация референсного генома в формате GFF; база данных мотивов ССТФ в формате MEME; классификация ТФ по структуре ДСД в формате TSV. Сырые данные картируются на референсный геном, который также используется для извлечения нуклеотидных последовательностей пиков ChIP-seq и подготовки геномных негативных последовательностей (Raditsa et al., 2024). Файлы геномной аннотации содержат полногеномную разметку генов, транскриптов и т. п. (Dyer et al., 2025). Эти данные необходимы для извлечения 5'-регуляторных районов генов (промоторов), которые, в свою очередь, необходимы для единообразного выбора порога мотивов разных моделей, см. далее. Для аннотации найденных *de novo* мотивов ССТФ используются коллекции известных

Таблица 1. Инструменты предварительной подготовки сырых данных ChIP-seq
Table 1. ChIP-seq raw data pre-processing tools

Инструмент	Функции/Назначение	Источник
Bowtie2	Картирование прочтений на референсный геном	Langmead et al., 2019
MACS2	Поиск пиков ChIP-seq и оценка их значимости	Zhang et al., 2008
bedtools	Извлечение пиков ChIP-seq по координатам, представленным в файле формата BED, для подготовки файла в формате FASTA с пиками ChIP-seq	Quinlan, Hall, 2010
AntiNoise	Экстракция негативных последовательностей ДНК из генома	Raditsa et al., 2024

Таблица 2. Дополнительные инструменты, необходимые для анализа данных ChIP-seq
Table 2. Additional tools required to analyze ChIP-seq data

Инструмент	Функции/Назначение	Источник
bedtools	Извлечение районов промоторов по координатам, представленным в файле формата BED, для подготовки файла в формате FASTA с районами промоторов	Quinlan, Hall, 2010
TFClass	Иерархическая классификация ТФ по структуре ДСД	Wingender et al., 2018
BSMotif	Кластеризация мотивов ССТФ по сходству на основе иерархической классификации TFClass	Levitsky et al., 2025b

мотивов (Bailey, Elkan, 1994) из баз данных мотивов сайтов связывания известных ТФ HOCOMOCO (человек и мышь) (Vorontsov et al., 2024) и JASPAR (основные таксоны эукариот) (Rauluseviciute et al., 2024). Мотивы, найденные с помощью *de novo* поиска, сравниваются с мотивами сайтов связывания известных ТФ из этих коллекций. Иерархическая классификация ТФ по типам ДСД используется для кластеризации и аннотации мотивов согласно структуре ДСД ТФ (Wingender et al., 2013, 2015, 2018).

Подготовка данных ChIP-seq. Используемые в этом блоке инструменты приведены в табл. 1. Данный блок конвейера отвечает за предварительную обработку сырых данных ChIP-seq и подготовку их к последующему поиску обогащенных мотивов. Основная задача блока заключается в преобразовании прочтений (.fastq), полученных на секвенаторе, в нуклеотидные последовательности ДНК (.fasta) геномных локусов (пиков).

На первом этапе обработки сырых данных ChIP-seq производится картирование прочтений на референсный геном с использованием инструмента Bowtie (Langmead et al., 2019; Carter et al., 2023), который широко применяется для работы с данными ChIP-seq (Oki et al., 2018; Kolmykov et al., 2021). Последующее определение пиков осуществляется с помощью инструмента MACS2 (Zhang et al., 2008), который идентифицирует области значимого обогащения прочтений относительно контрольных образцов; существуют аналоги MACS2, но данный инструмент показывает более высокое качество результатов для ChIP-seq экспериментов по ТФ (Malone et al., 2011; Lavrekha et al., 2022). Из идентифицированных пиков ChIP-seq извлекаются нуклеотидные последовательности с помощью инструмента bedtools (Quinlan, Hall, 2010).

Задачу подготовки выборки негативных последовательностей ДНК решает инструмент AntiNoise (Raditsa et al., 2024). В отличие от популярных подходов, генерирующих синтетические негативные последовательности по простым моделям нуклеотидного состава, AntiNoise извлекает геномные последовательности, точно соответствующие исходным пикам по длине и G/C-составу (Raditsa et al., 2024). Геномный подход AntiNoise к подготовке негативных выборок обеспечивает снижение доли ложноположительных результатов за счет сохранения геном-специфических особенностей состава нуклеотидов (Raditsa et al., 2024).

Подготовка общих данных. Используемые в этом блоке инструменты приведены в табл. 2. Этот блок конвейера обеспечивает подготовку вспомогательных данных, необходимых для контекстной интерпретации результатов анализа мотивов. Основная функция блока заключается в создании структурированной базы данных для сравнения и аннотации найденных мотивов ССТФ.

Из референсного генома извлекаются последовательности промоторных районов генов (обычно от 500 до 5000 п.н. выше сайта старта транскрипции) для последующего анализа. Промоторные области используются как для оценки частоты встречаемости мотивов, так и для функциональной аннотации мотивов ССТФ через анализ генов.

На основе базы данных TFClass и используемой базы мотивов с помощью инструмента BSMotif создается гибридная иерархическая система классификации ТФ, которая учитывает как структуру ДСД ТФ согласно TFClass, так и сходство мотивов ССТФ. Определяется полный набор ветвей ТФ, каждая из которых может включать один класс ТФ, или одно или несколько семейств класса, или одно или несколько подсемейств, или один или несколько ТФ, и в каждом случае также все более низкие уровни иерархии, имеющие схожие мотивы. Для каждой ветви большинство возможных пар ТФ имеет значимо сходные мотивы ССТФ. Таким образом, создается связь между структурной классификацией ТФ и вариабельностью мотивов их сайтов связывания. Это позволяет эффективнее аннотировать и интерпретировать результаты как *de novo* поиска, так и других инструментов, связанных с поиском обогащенных мотивов, поскольку ко-

Таблица 3. Инструменты анализа данных ChIP-seq
Table 3. ChIP-seq data analysis tools

Инструмент	Функции/Назначение	Источник
MultiDeNA	Единообразное применение и интеграция разных моделей мотивов для *de novo* поиска	Tsukanov et al., 2022
STREME	Поиск мотивов на основе традиционной модели PWM	Bailey, 2021
BaMM	Поиск мотивов на основе альтернативной модели, байесовские марковские модели с учетом зависимостей позиций	Ge et al., 2021
InMoDe	Поиск мотивов на основе альтернативной модели, интерполированные марковские модели на основе контекстных деревьев	Eggeling et al., 2017
SiteGA	Поиск мотивов на основе альтернативной модели, генетический алгоритм для набора локально-позиционированных динуклеотидов	Tsukanov et al., 2022
MCOT	Анализ совместной встречаемости в парах мотивов ССТФ	Levitsky et al., 2019, 2020, 2025a
MetArea	Анализ взаимоисключающей встречаемости в парах мотивов ССТФ	Levitsky et al., 2024
TomTom	Оценка сходства мотивов модели PWM	Gupta et al., 2007
clusterProfiler	Анализ обогащения терминов генной онтологии для пиков с предсказанными ССТФ	Wu et al., 2021

личество уникальных мотивов меньше, чем количество ТФ (Castro-Mondragon et al., 2017; Pratt et al., 2022; Rauluseviciute et al., 2024).

Анализ данных ChIP-seq с применением мотивов нескольких моделей. Используемые в этом блоке инструменты приведены в табл. 3. Центральный блок конвейера реализует интегрированный подход к поиску и анализу мотивов ССТФ нескольких моделей. Блок обеспечивает применение различных моделей *de novo* поиска мотивов с последующей процедурой перекрестной проверки для оценки точности распознавания и интеграцией результатов.

Центральный компонент MultiDeNA запускает поиск мотивов с использованием четырех различных моделей мотива ССТФ: PWM (STREME), BaMM, InMoDe, SiteGA. Для каждой модели применяют процедуру перекрестной проверки, чтобы подобрать оптимальные параметры модели мотива и оценить точность распознавания по метрикам auROC и auPRC (Lever, 2016). Используется схема двукратной перекрестной проверки с разделением всей выборки пиков на обучающую и контрольную выборки, оценка точности определяется на заранее подготовленной с помощью инструмента AntiNoise (Raditsa et al., 2024) негативной выборке.

Для всех моделей единообразно определяются пороги функций распознавания для последующего предсказания ССТФ в последовательностях. Пороги устанавливаются исходя из заданного уровня ложноположительных предсказаний, определяемых по частоте встречаемости мотива в полногеномной выборке промоторов (Levitsky et al., 2019; Tsukanov et al., 2022). Такая процедура позволяет единообразно установить пороги для методологически разных моделей.

С использованием обученных при оптимальных параметрах мотивов разных моделей и выбранных пороговых значений производится распознавание ССТФ в последовательностях ChIP-seq пиков. Каждое предсказание характеризуется позицией в последовательности, цепью ДНК и значением функции распознавания, результаты для мотива каждой модели записываются в формате bed. Использование разных моделей мотива в MultiDeNA позволяет выявлять различные структурные типы сайтов связывания, которые могут быть пропущены при использовании одной модели PWM (Tsukanov et al., 2022).

Все пики ChIP-seq классифицируются на основе наличия предсказаний ССТФ: пики с предсказаниями только одной модели, пики с перекрывающимися предсказаниями нескольких моделей, пики без предсказаний (Tsukanov et al., 2021). Такая классификация позволяет оценить вклады мотивов разных моделей в общее распознание сайтов связывания целевого ТФ в данных ChIP-seq. Далее определяется набор генов, промоторы которых содержат предсказанные ССТФ по каждой из моделей мотива. Это делается путем пересечения разметки промоторов генов всего генома и списков пиков, содержащих предсказанные ССТФ, затем с помощью инструмента clusterProfiler (Wu et al., 2021) проводится анализ терминов генной онтологии (ГО). Эти данные используются для интерпретации биологической функции генов, в промоторах которых были предсказаны сайты мотивом определенной модели, и выявления ассоциации этих генов с конкретными терминами генной онтологии.

Инструменты MCOT (Levitsky et al., 2019) и MetArea (Levitsky et al., 2024) применяют для анализа пар мотивов ССТФ. Это может объяснить механизм связывания в пиках ChIP-seq, в которых: целевой ТФ связывается в виде димера или тетрамера (Amoutzias, 2008; MCOT, Levitsky et al., 2025a); есть мотивы сайтов связывания партнерских ТФ, более консервативные, чем мотивы сайтов связывания целевого ТФ (MCOT, Levitsky et al., 2020); нет мотивов сайтов связывания целевых ТФ, есть мотивы сайтов связывания партнерских ТФ (MetArea, Levitsky et al., 2024).

Результаты и обсуждение

Предложенный конвейер MultiDeCA был применен для анализа ChIP-seq данных по ТФ JunD. Данные эксперимента ChIP-seq получены на макрофагах костного мозга мыши, обработанных 100 нг/мл липидом А в течение 30 мин (GEO ID

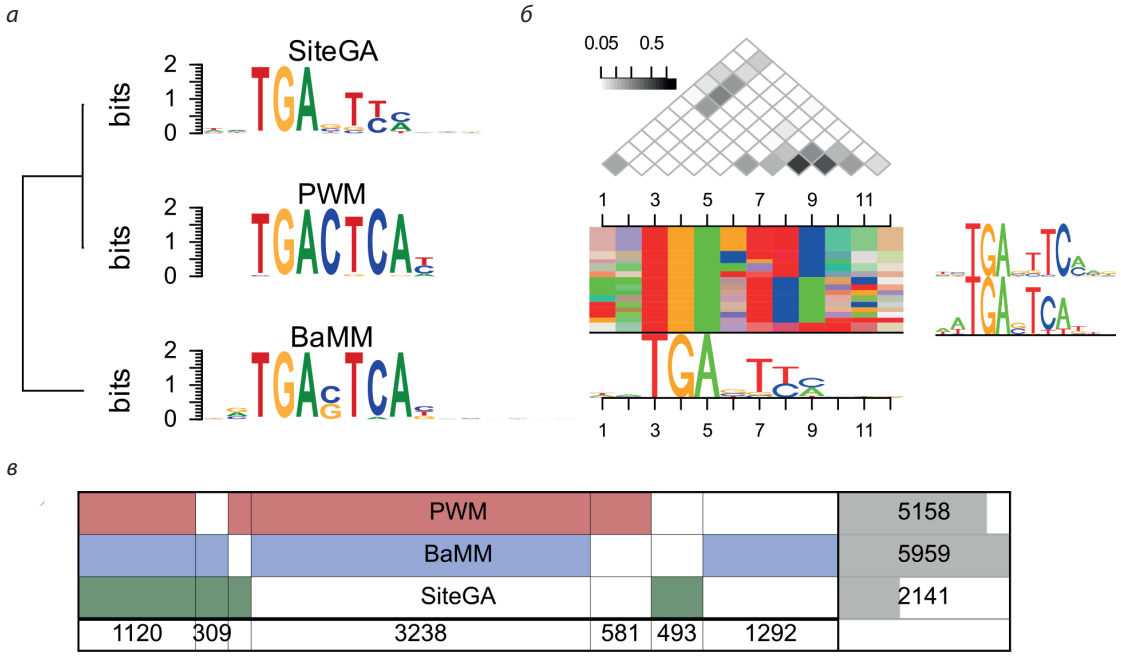

Рис. 2. Результаты работы MultiDeNA:

а – лого-диаграммы для моделей мотива PWM, BaMM и SiteGA; *б* – DepLogo (Grau et al., 2019) включает лого-диаграммы (справа и внизу), цветные полосы в центре и треугольную матрицу сверху. Цветные полосы отражают частоты нуклеотидов в позициях мотива. Треугольная матрица отражает зависимости разных позиций с помощью метрики взаимной информации, показано оттенками серого. В правой части – две лого-диаграммы, отражающие два структурных типа ССТФ, предсказанных моделью SiteGA, по длине спейсера в один или два нуклеотида; *в* – число пиков с предсказанными ССТФ моделей мотивов отдельно (PWM/BaMM/SiteGA – красный/синий/зеленый) и вместе и любым сочетанием их пар, представленные диаграммой superVenn (Indukaev, 2024). Горизонтальная/вертикальная линии внизу/справа диаграммы указывают число пиков (показаны фракции 300 пиков и более).

Fig. 2. MultiDeNA tool results.

a – logo for motif models PWM, BaMM and SiteGA; *b* – DepLogo (Grau et al., 2019) includes logo diagrams (right and bottom), colored strips in the center, and a triangular matrix on top. The colored strips show the nucleotide frequencies at the motif positions. The triangular matrix reflects the dependencies of different positions using the metric mutual information reflected in shades of gray. The right side displays two logos reflecting the two structural types of TFBS predicted by the SiteGA model, with a spacer length of 1 or 2 nt; *c* – the number of peaks with predicted TFBSs by the motif models separately (PWM/BaMM/SiteGA – red/blue/green) and together, and by any combination of their pairs, represented by the superVenn diagram (Indukaev, 2024). The horizontal/vertical lines at the bottom/right of the diagram indicate the number of peaks (fractions of 300 or more peaks are shown).

GSM2663855). Исходные данные были взяты из базы GTRD (ID PEAKS040980) в предобработанном виде, т. е. они были представлены в виде пиков, идентифицированных с помощью инструмента MACS2. Всего пиков в наборе 107 026, из них только 1000 пиков лучшего качества (наибольшие значения отрицательного логарифма значимости – $\log_{10}$(*p*-value)) были отобраны для *de novo* поиска мотивов. Далее выполняли следующие шаги анализа: 1) с помощью инструмента bedtools извлечены нуклеотидные последовательности из полного генома; 2) с помощью инструмента AntiNoise получена негативная выборка для *de novo* поиска; 3) инструмент MultiDeNA осуществил подбор параметров моделей мотива PWM (Streme), BaMM и SiteGA с помощью перекрестной проверки и *de novo* поиск мотивов этих моделей; 4) инструмент MultiDeNA сравнил готовые обогащенные мотивы с мотивами сайтов связывания известных ТФ из базы данных HOCOMOCO с помощью инструмента TomTom, чтобы оценить корректность предсказанных ССТФ и определить возможные ТФ-регуляторы; 5) инструмент MultiDeNA выполнил предсказание ССТФ на заданном пороге, в нашем случае поиск осуществляли на всем объеме выборки данных; 6) инструмент MultiDeNA для пиков, содержащих предсказанные ССТФ, оценил вклад каждой модели и провел аннотацию, т. е. пики с ССТФ накладывались на промоторы генов, и оценил обогащение терминов GO для фракций пиков с разными сочетаниями присутствия предсказанных сайтов разных моделей мотивов; 7) с помощью инструмента MCOT выявили композиционные элементы (КЭ), что указало на потенциальные партнерские ТФ для целевого ТФ JunD.

Полученные мотивы трех моделей PWM/BaMM/SiteGA в виде лого-диаграмм показаны на рис. 2, *а*. Все три мотива имели значимое сходство (TomTom *p*-value < 0.0001) с мотивами сайтов связывания, относящимися к той же ветви ТФ, что и целевой ТФ JunD согласно полученной с помощью инструмента BSMotif кластеризации мотивов из базы данных HOCOMOCO. Можно отметить, что для мотивов всех моделей совпадают первые три консервативных нуклеотида (TGA, см. рис. 2, *а*). В четвертой позиции (относительно первого T) для SiteGA наблюдается отсутствие консервативности, т. е. может быть любой нуклеотид, для PWM консервативный C, а для BaMM C/G. С пятой по седьмую позицию у мотивов модели PWM и BaMM наблюдается более высокий уровень консервативности нуклеотидов, чем у мотива модели SiteGA. Дополнительный анализ мотива модели SiteGA

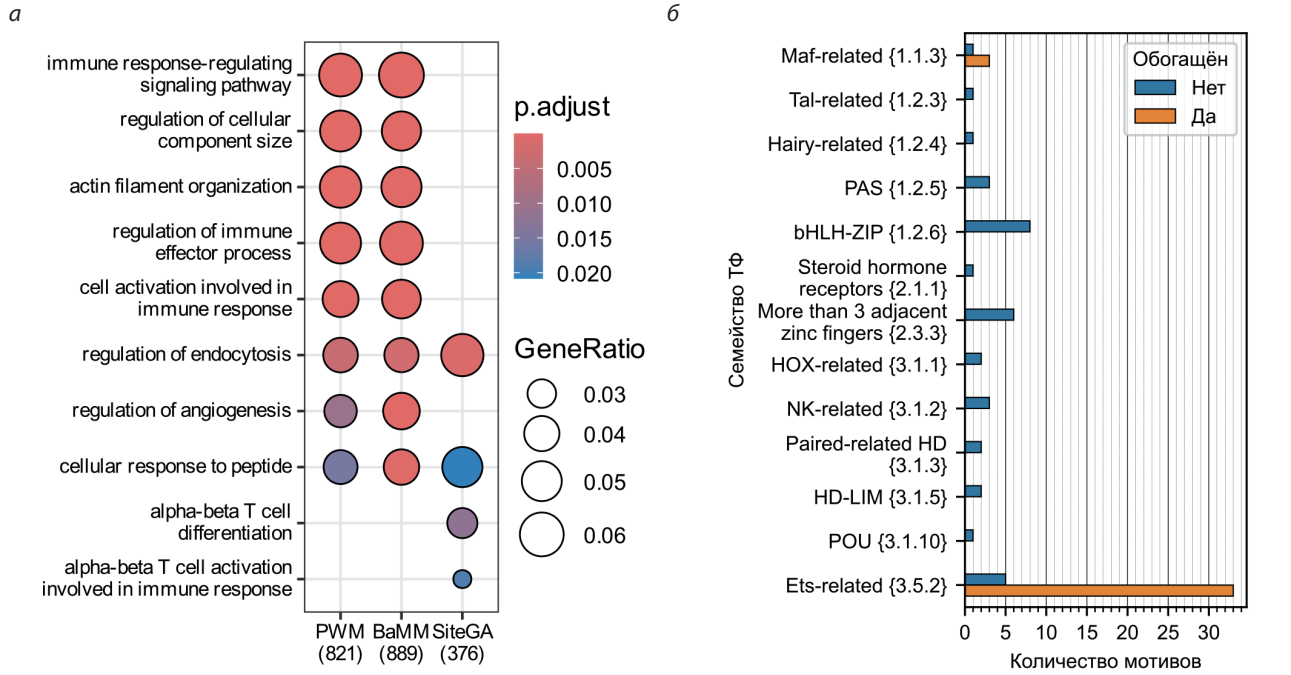

Рис. 3. Результаты работы инструментов clusterProfiler и MCOT:

а – анализ терминов ГО. Значимо обогащенные термины ГО для набора данных ChIP-seq для ТФ JunD (GTRD ID PEAKS040980, GEO ID GSM2663855), полученные по результатам предсказания сайтов мотивов моделей PWM, BaMM и SiteGA в промоторы генов. На оси *X* обозначены модели мотивов. На оси *Y* перечислены термины ГО. Размер кружка отражает долю генов (GeneRatio), ассоциированных с термином ГО для анализируемого списка генов. Цвет отображает значимость обогащения термина ГО (скорректированное значение *p*-value, p.adjust); *б* – количество мотивов ССТФ партнерских ТФ, найденных с использованием MCOT и классифицированных по семействам из базы данных TFClass (Wingender et al., 2018). Цифры в фигурных скобках обозначают иерархические уровни согласно TFClass, например, семейство Maf-related factors {1.1.3} относится к классу Basic leucine zipper factors (bZIP) {1.1} и суперклассу Basic domains {1}. Красным/синим цветом обозначено число мотивов семейства, для которых есть/нет значимое обогащение в пиках ChIP-seq по сравнению с негативной выборкой последовательностей ДНК от инструмента AntiNoise, оценка обогащения получена с помощью инструмента AME (McLeay, Bailey, 2010) для коллекции мотивов HOCOMOCO версии 12, порог значимости с поправкой на множественные сравнения, adjusted *p*-value < 0.01.

Fig. 3. Results of ClusterProfiler and MCOT tools.

a – analysis of GO terms. The significantly enriched GO terms for JunD TF ChIP-seq dataset (GTRD ID PEAKS040980, GEO ID GSM2663855) derived from recognition of sites by the motif of PWM, BaMM, and SiteGA models in gene promoters. The *X*-axis indicates the motif models. The *Y*-axis lists the GO terms. The size of the circle reflects the proportion of genes (GeneRatio) associated with the GO term for the analyzed gene list. Color displays the enrichment significance of the GO term (adjusted *p*-value, p.adjust); *b* – number of TFBS motifs of partner TFs found by MCOT and classified by families from the TFClass database (Wingender et al., 2018). Numbers in curly brackets indicate hierarchical levels according to the TFClass, for example, the Maf-related factors {1.1.3} family belongs to the Basic leucine zipper factors (bZIP) {1.1} class and the Basic domains {1} superclass. Red/blue indicates the number of TFBS motifs per family for which is present/absent the significant enrichment in the ChIP-seq peaks compared to the negative sequence set from AntiNoise tool, enrichment estimates obtained by the AME tool for the HOCOMOCO motif collection, significance threshold adjusted for multiple comparisons, adjusted *p*-value < 0.01.

с помощью инструмента DepLogo (Grau et al., 2019) (см. рис. 2, *б*) показал, что сниженный уровень консервативности нуклеотидов на правом фланге мотива связан с тем, что модель SiteGA предсказывает ССТФ, где между левым флангом (TGA) и правым (TCA) может быть спейсер длины 1 или 2 п.н. Обнаруженная особенность отражает то, что ТФ JunD относится к классу Basic leucine zipper factors (bZIP), а ТФ этого класса связываются с ДНК только как димеры (Amoutzias et al., 2008), при этом два полусайта всегда лежат в одной цепи, а спейсер между ними допускает лишь малую вариацию (Nagy G., Nagy L., 2020).

С помощью инструмента MultiDeNA в 10 000 лучших пиков ChIP-seq были предсказаны ССТФ с помощью мотивов трех моделей на единообразно подобранных порогах распознавания (Tsukanov et al., 2022), которые соответствуют ошибке перепредсказания равной 10^{-4}. Количество пиков, содержащих предсказания ССТФ мотивами разных моделей, показано на рис. 2, *в*. Модель PWM выявила ССТФ в 5158 пиках, а модели BaMM и SiteGA к этому количеству независимо добавили еще 493 и 1292 соответственно и еще 309, где предсказали BaMM и SiteGA одновременно, что демонстрирует существенный прирост в количестве пиков, содержащих ССТФ (общий прирост составил 2094). Можно предположить, что количество прямых взаимодействий ТФ JunD с геномной ДНК может быть больше, чем предсказывает мотив модели PWM. Следует отметить, что ТФ JunD входит в состав димера (ТФ AP-1), с которым связываются гомо- и гетеродимеры, состоящие из ТФ Jun (v-Jun, c-Jun, JunB и JunD), Fos (v-Fos, c-Fos, FosB, Fra1 и Fra2), ATF (ATF2, ATF3/LRF1, B-ATF, JDP1 и JDP2) и MAF (c-Maf, MafB, MafA, MafG/F/K и Nrl) (Chinenov, Kerppola, 2001). Из-за этого может наблюдаться структурное разнообразие мотивов ССТФ для JunD, которые находят модели BaMM и SiteGA, но не находит модель PWM.

Пики с предсказанными ССТФ пересекали с промоторами (–2000, +300 относительно сайта старта транскрипции) генов мыши, кодирующих белки (версия генома GRCm38.102).

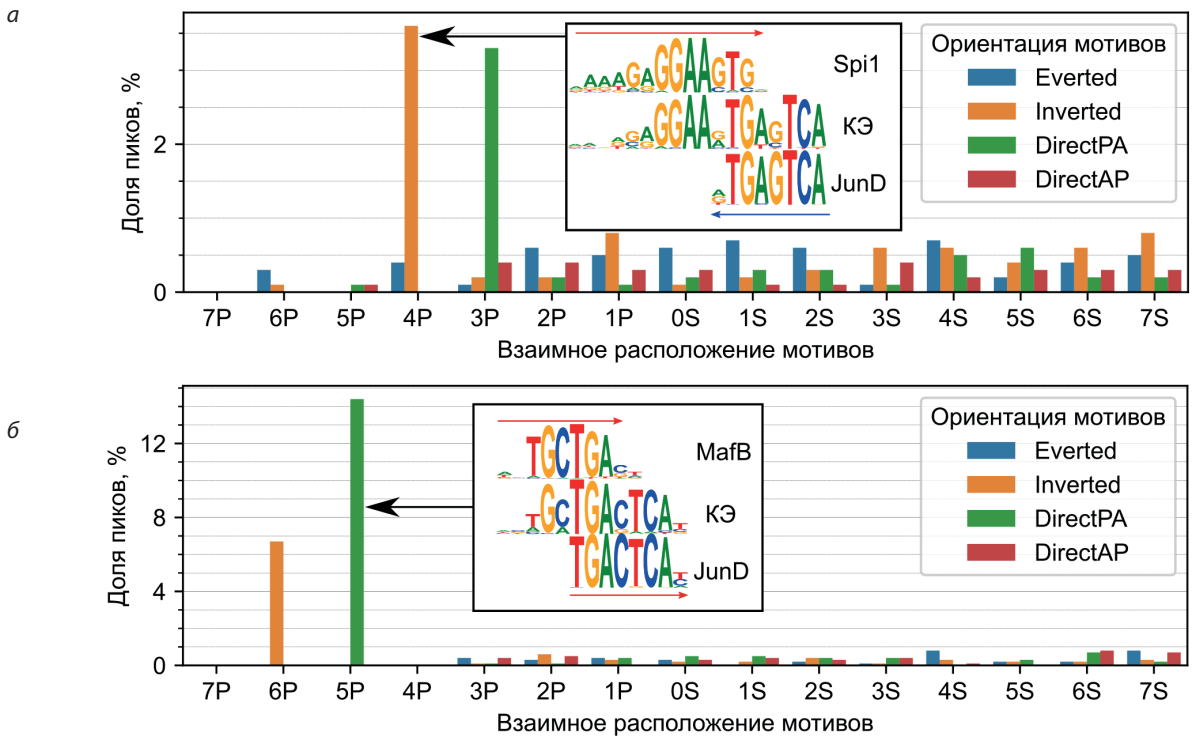

Рис. 4. Примеры предсказанных КЭ инструментом MCOT для пар ТФ:

а – JunD/Spi1; *б* – JunD/MafB. Мотивы SPI1.H12CORE.0.P.B и MAFB.H12CORE.1.P.B из коллекции HOCOMOCO версии 12, мотив JunD – из результатов *de novo* поиска для модели PWM. Структура КЭ определяется длиной спейсера или перекрывания сайтов и их взаимной ориентацией (Levitsky et al., 2019), выравнивание лого-диаграмм и красная/синяя стрелки показывают расположение сайтов у самого частого варианта КЭ (черные стрелки). Ось *X* обозначает взаимное расположение двух сайтов, диапазоны частичных перекрытий/спейсеров имеют суффиксы P/S. Ось *Y* обозначает долю пиков, содержащих КЭ с определенным взаимным расположением и ориентацией сайтов. Четыре типа ориентаций взяты согласно (Levitsky et al., 2019). Два сайта в одной цепи: Direct PA/Direct AP: сначала сайт партнерского/целевого ТФ, а потом целевого/партнерского ТФ (в обоих случаях 5′→3′, 5′→3′). Два сайта в разных цепях: Inverted (хвост к хвосту, 5′→3′, 3′→5′) и Everted (голова к голове, 3′→5′, 5′→3′).

Fig. 4. Examples of predicted CEs by the MCOT tool for TF pairs.

a – JunD/Spi1; *b* – JunD/MafB. SPI1.H12CORE.0.P.B and MAFB.H12CORE.1.P.B motifs from the HOCOMOCO version 12 collection, JunD motif is taken from the *de novo* search result for the PWM model. The structure of CEs is determined by the spacer or overlap length, and the mutual orientation of two sites (Levitsky et al., 2019), the alignment of logo s and red/blue arrows show the location of sites at the most frequent CE variant (black arrows). The *X*-axis denotes the mutual location of two sites, the ranges of partial overlaps/spacers are suffixed with P/S. The *Y*-axis shows the fraction of peaks with CEs of a certain mutual location and orientation of sites. Four orientation types were selected according to (Levitsky et al., 2019). Two sites are in the same strand: Direct PA/Direct AP mean the first site of the partner/target TF, and then the second of the target/partner TF (both cases imply 5′→3′, 5′→3′). Two sites in the different strands: Inverted (tail to tail, 5′→3′, 3′→5′) and Everted (head to head, 3′→5′, 5′→3′).

Далее с помощью инструмента clusterProfiler оценивали обогащение терминов ГО для каждой модели мотива. Всего было выявлено 56, 136 и 12 терминов ГО для PWM, BaMM и SiteGA соответственно. Некоторые термины ГО могут быть близкими по семантике, поэтому далее с помощью встроенной функции инструмента clusterProfiler были удалены избыточные термины. Количество терминов сократилось до 31, 68 и 4 для PWM, BaMM и SiteGA соответственно. Первые десять наиболее обогащенных терминов ГО показаны на рис. 3, *а*. Большинство терминов ГО относится к биологическим процессам иммунного ответа, что указывает на роль JunD в транскрипционном контроле воспалительных реакций. Кроме того, выявлены термины, связанные с организацией цитоскелета (actin filament organization) и эндоцитоза (regulation of endocytosis), что подчеркивает мультифункциональность JunD в адаптации макрофагов к стимуляции липидом А. Термины, связанные с Т-клеточной дифференцировкой, обнаруженные только у мотива модели SiteGA, указывают на потенциальные перекрестные механизмы регуляции между врожденным и адаптивным иммунитетом. В целом этот результат подтверждает результаты анализа других ТФ (Levitsky et al., 2016; Tsukanov et al., 2022), в которых установлено, что модели SiteGA и PWM находят сайты в промоторах генов, описываемых не полностью перекрывающимися терминами генной онтологии.

Часто в регуляции транскрипции генов участвует несколько ТФ. Чтобы установить партнерские ТФ для целевого ТФ JunD, применили инструмент MCOT. Поиск ТФ-партнеров делали с использованием базы мотивов ССТФ HOCOMOCO, а в качестве модели мотива сайтов связывания ТФ JunD взяли PWM. Значимых ТФ партнеров выбирали с помощью MCOT по следующим критериям: мотив ССТФ партнера значимо не похож на мотив ССТФ JunD, нет значимости КЭ для полного перекрывания мотивов, есть значимость КЭ с частичным перекрыванием или спейсером.

Всего было выявлено 72 мотива для 56 партнерских ТФ, относящихся к 13 разным семействам и 6 классам ТФ. Далее с помощью инструмента AME оценили, насколько мотивы

сайтов связывания партнерских ТФ обогащены в пиках ChIP-seq. В итоге обогащены (adjusted *p*-value < 0.01) были только 36 мотивов для 23 партнерских ТФ, относящихся к двум семействам (см. рис. 3, *б*).

Примечательно, что большинство партнерских ТФ относится к семейству Ets-related {3.5.2} (см. рис. 3, *б*). Транскрипционные факторы из этого семейства играют важную роль в контроле дифференцировки, метаболизма и воспалительных функций макрофагов (Yang et al., 2024), что согласуется с условиями эксперимента и результатами анализа генной онтологии. При этом в более ранних исследованиях уже было показано, что есть белок-белковое взаимодействие между ТФ JunD и ТФ из семейства Ets-related {3.5.2} (Chinenov, Kerppola, 2001; Oughtred et al., 2021). Партнерским ТФ из семейства Ets-related {3.5.2} может быть также известный ТФ SPI1 (Pu. 1, рис. 4, *а*), играющий важную роль в макрофагах костного мозга (Heinz et al., 2010). Можно предположить формирование *in vivo* комплекса из трех ТФ: два из них ТФ AP-1 (димер ТФ JunD с родственным ТФ из семейств Jun, Maf или Fos), а третий ТФ SPI1, известный своей пионерной активностью (Levitsky et al., 2025a).

Следующее семейство Maf-related, к нему относится ТФ MafB, мотив которого также оказался обогащенным. MafB участвует в моноцитарной дифференцировке и поддержании идентичности макрофагов (Ventura et al., 2025). Как показано выше, этот ТФ также может участвовать в димеризации с другими ТФ (Jun, Fos), формируя ТФ AP-1 (Chinenov, Kerppola, 2001). Полученные результаты по партнерским ТФ согласуются с результатами анализа терминов ГО, а также демонстрируют широкий спектр ТФ партнеров, которые участвуют в регуляции транскрипции совместно с JunD.

Заключение

Представленный конвейер MultiDeCA объединяет специализированные биоинформатические инструменты для комплексного анализа ChIP-seq данных и поиска мотивов сайтов связывания транскрипционных факторов. Конвейер обеспечивает полный цикл анализа от исходных данных до биологической интерпретации результатов, включая подготовку негативных последовательностей, единообразные для разных моделей мотива процедуры перекрестной проверки и обучения моделей мотива *de novo*, их совместное применение, структурную классификацию мотивов, относящихся к отдельным известным ТФ, анализ совместной и взаимоисключающей встречаемостей в парах мотивов.

Пример применения конвейера MultiDeCA к анализу ChIP-seq данных по ТФ JunD продемонстрировал эффективность конвейера на практике. Использование трех различных моделей мотивов позволило существенно расширить набор предсказанных сайтов связывания по сравнению со стандартным подходом на основе только модели PWM: модели BaMM и SiteGA дополнительно выявили 2094 из всех 10 000 пиков, содержащих потенциальные ССТФ (20.94 %, добавленные к 51.58 % от модели PWM). Это указывает на то, что более сложные модели, чем традиционная PWM, способны учитывать структурные особенности сайтов связывания, которые не улавливаются классическими позиционно-весовыми матрицами. Последующий анализ терминов ГО и поиск партнерских ТФ продемонстрировали биологическую осмысленность полученных результатов. Выявленные термины ГО соответствуют известной роли JunD в регуляции иммунного ответа и воспалительных процессов в макрофагах. Идентификация 23 партнерских ТФ, преимущественно из семейств Ets-related и Maf-related, подтверждает участие ТФ JunD в сложных регуляторных сетях, контролирующих дифференцировку и функции макрофагов. Согласованность результатов, полученных разными методами анализа, свидетельствует о надежности предложенного конвейера и его применимости для изучения механизмов транскрипционной регуляции.

Список литературы / References

Amoutzias G.D., Robertson D.L., Van de Peer Y., Oliver S.G. Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem Sci*. 2008;33(5):220-229. doi 10.1016/j.tibs.2008.02.002

Bailey T.L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*. 2021;37(18):2834-2840. doi 10.1093/bioinformatics/btab203

Bailey T.L., Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, 1994;2:28-36

Carter J.L., Stevens H., Ridge P.G., Johnson S.M. Short sequence aligner benchmarking for chromatin research. *Int J Mol Sci*. 2023;24(18):14074. doi 10.3390/ijms241814074

Castro-Mondragon J.A., Jaeger S., Thieffry D., Thomas-Chollier M., van Helden J. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res*. 2017;45(13):e119. doi 10.1093/nar/gkx314

Chinenov Y., Kerppola T.K. Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity. *Oncogene*. 2001;20(19):2438-2452. doi 10.1038/sj.onc.1204385

Dyer S.C., Austine-Orimoloye O., Azov A.G., Barba M., Barnes I., Barrera-Enriquez V.P., Becker A., … Finn R.D., Freeberg M.A., Harrison P.W., Martin F.J., Yates A.D. Ensembl 2025. *Nucleic Acids Res*. 2025;53(D1):D948-D957. doi 10.1093/nar/gkae1071

Eggeling R., Grosse I., Grau J. InMoDe: Tools for learning and visualizing intra-motif dependencies of DNA binding sites. *Bioinformatics*. 2017;33(4):580-582. doi 10.1093/bioinformatics/btw689

Farnham P.J. Insights from genomic profiling of transcription factors. *Nat Rev Genet*. 2009;10(9):605-616. doi 10.1038/nrg2636

Furey T.S. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet*. 2012;13(12):840-852. doi 10.1038/nrg3306

Ge W., Meier M., Roth C., Söding J. Bayesian Markov models improve the prediction of binding motifs beyond first order. *NAR Genom Bioinform*. 2021;3(2):lqab026. doi 10.1093/nargab/lqab026

Grau J., Nettling M., Keilwagen J. DepLogo: visualizing sequence dependencies in R. *Bioinformatics*. 2019;35(22):4812-4814. doi 10.1093/bioinformatics/btz507

Gupta S., Stamatoyannopoulos J.A., Bailey T.L., Noble W. Quantifying similarity between motifs. *Genome Biol*. 2007;8(2):R24. doi 10.1186/gb-2007-8-2-r24

Heinz S., Benner C., Spann N., Bertolino E., Lin Y.C., Laslo P., Cheng J.X., Murre C., Singh H., Glass C.K. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010;38(4):576-589. doi 10.1016/j.molcel.2010.05.004

Indukaev F. gecko984/supervenn: v0.5.0. *Zenodo*. 2024. doi 10.5281/zenodo.11395173

Kel O.V., Romaschenko A.G., Kel A.E., Wingender E., Kolchanov N.A. A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res*. 1995;23(20):4097-4103. doi 10.1093/nar/23.20.4097

Kolmykov S., Yevshin I., Kulyashov M., Sharipov R., Kondrakhin Y., Makeev V.J., Kulakovskiy I.V., Kel A., Kolpakov F. GTRD: an integrated

view of transcription regulation. *Nucleic Acids Res*. 2021;49(D1): D104-D111. doi 10.1093/nar/gkaa1057
Lambert S.A., Jolma A., Campitelli L.F., Das P.K., Yin Y., Albu M., Chen X., Taipale J., Hughes T.R., Weirauch M.T. The human transcription factors. *Cell*. 2018;172(4):650-665. doi 10.1016/j.cell.2018.01.029
Langmead B., Wilks C., Antonescu V., Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*. 2019;35(3):421-432. doi 10.1093/bioinformatics/bty648
Lavrekha V.V., Levitsky V.G., Tsukanov A.V., Bogomolov A.G., Grigorovich D.A., Omelyanchuk N., Ubogoeva E.V., Zemlyanskaya E.V., Mironova V. CisCross: A gene list enrichment analysis to predict upstream regulators in *Arabidopsis thaliana*. *Front Plant Sci*. 2022;13:942710. doi 10.3389/fpls.2022.942710
Lever J., Krzywinski M., Altman N. Classification evaluation. *Nat Methods*. 2016;13:603-604. doi 10.1038/nmeth.3945
Levitsky V.G., Ignatieva E.V., Ananko E.A., Turnaev I.I., Merkulova T.I., Kolchanov N.A., Hodgman T.C.T. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. *BMC Bioinformatics*. 2007;8(1):481. doi 10.1186/1471-2105-8-481
Levitsky V.G., Oshchepkov D.Y., Klimova N.V., Ignatieva E.V., Vasiliev G.V., Merkulov V.M., Merkulova T.I. Hidden heterogeneity of transcription factor binding sites: A case study of SF-1. *Comput Biol Chem*. 2016;64:19-32. doi 10.1016/j.compbiolchem.2016.04.008
Levitsky V.G., Zemlyanskaya E.V., Oshchepkov D.Yu., Podkolodnaya O.A., Ignatieva E.V., Grosse I., Mironova V.V., Merkulova T.I. A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. *Nucleic Acids Res*. 2019;47(21): e139. doi 10.1093/nar/gkz800
Levitsky V.G., Oshchepkov D.Yu., Zemlyanskaya E.V., Merkulova T.I. Asymmetric conservation within pairs of co-occurred motifs mediates weak direct binding of transcription factors in ChIP-Seq data. *Int J Mol Sci*. 2020;21(17):6023. doi 10.3390/IJMS21176023
Levitsky V.G., Tsukanov A.V., Merkulova T.I. MetArea: a software package for analysis of the mutually exclusive occurrence in pairs of motifs of transcription factor binding sites based on ChIP-seq data. *Vavilov J Genet Breed*. 2024;28(8):822-833. doi 10.18699/vjgb-24-90
Levitsky V.G., Raditsa V.V., Tsukanov A.V., Mukhin A.M., Zhimulev I.F., Merkulova T.I. Asymmetry of motif conservation within their homotypic pairs distinguishes DNA-binding domains of target transcription factors in ChIP-Seq data. *Int J Mol Sci*. 2025a;26(1):386. doi 10.3390/ijms26010386
Levitsky V.G., Vatolina T.Yu., Raditsa V.V. Linking hierarchical classification of transcription factors by the structure of their DNA-binding domains to the variability of their binding site motifs. *Vavilov J Genet Breed*. 2025b;29(7):925-939. doi 10.18699/vjgb-25-99
Lihu A., Holban Ş. A review of ensemble methods for *de novo* motif discovery in ChIP-Seq data. *Brief Bioinform*. 2015;16(6):964-973. doi 10.1093/bib/bbv022
Malone B.M., Tan F., Bridges S.M., Peng Z. Comparison of four ChIP-Seq analytical algorithms using rice endosperm H3K27 trimethylation profiling data. *PLoS One*. 2011;6(9):e25260. doi 10.1371/journal.pone.0025260
McLeay R. C., Bailey, T. L. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, 2010;11(1):165. https://doi.org/10.1186/1471-2105-11-165
Nagy G., Nagy L. Motif grammar: The basis of the language of gene expression. *Comput Struct Biotechnol J*. 2020;18:2026-2032. doi 10.1016/j.csbj.2020.07.007
Oki S., Ohta T., Shioi G., Hatanaka H., Ogasawara O., Okuda Y., Kawaji H., Nakaki R., Sese J., Meno C. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep*. 2018;19(12): e46255. doi 10.15252/embr.201846255
Oughtred R., Rust J., Chang C., Breitkreutz B.-J., Stark C., Willems A., Boucher L., … Dolma S., Coulombe-Huntington J., Chatr-Aryamontri A., Dolinski K., Tyers M. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci*. 2021;30(1):187-200. doi 10.1002/pro.3978
Pratt H.E., Andrews G.R., Phalke N., Huey J.D., Purcaro M.J., van der Velde A., Moore J.E., Weng Z. Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites. *Nucleic Acids Res*. 2022;50(D1):D141-D149. doi 10.1093/nar/gkab1039
Quinlan A.R., Hall I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842. doi 10.1093/bioinformatics/btq033
Raditsa V.V., Tsukanov A.V., Bogomolov A.G., Levitsky V.G. Genomic background sequences systematically outperform synthetic ones in *de novo* motif discovery for ChIP-seq data. *NAR Genom Bioinform*. 2024;6(3):lqae090. doi 10.1093/nargab/lqae090
Rauluseviciute I., Riudavets-Puig R., Blanc-Mathieu R., Castro-Mondragon J.A., Ferenc K., Kumar V., Lemma R.B., … Lenhard B., Sandelin A., Wasserman W.W., Parcy F., Mathelier A. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2024;52(D1):D174-D182. doi 10.1093/nar/gkad1059
Siebert M., Söding J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*. 2016;44:6055-6069. https://doi.org/10.1093/nar/gkw521
Simcha D., Price N.D., Geman D. The limits of *de novo* DNA motif discovery. *PloS One*. 2012;7(11):e47836. doi 10.1371/journal.pone.0047836
Stormo G.D. DNA binding sites: representation and discovery. *Bioinformatics*. 2000;16(1):16-23. doi 10.1093/bioinformatics/16.1.16
Tsukanov A.V., Levitsky V.G., Merkulova T.I. Application of alternative *de novo* motif recognition models for analysis of structural heterogeneity of transcription factor binding sites: a case study of FOXA2 binding sites. *Vavilov J Genet Breed*. 2021;25(1):7-17. doi 10.18699/VJ21.002
Tsukanov A.V., Mironova V.V., Levitsky V.G. Motif models proposing independent and interdependent impacts of nucleotides are related to high and low affinity transcription factor binding sites in Arabidopsis. *Front Plant Sci*. 2022;13:2637. doi 10.3389/FPLS.2022.938545
Ventura A.B., Loconte T., Negri A., Viggiano L., Fiermonte G., Ciavarella S., Guarini A., Volpe G. MAFB: a key regulator of myeloid commitment involved in hematological diseases. *Cell Death Discovery*. 2025;11(1):276. doi 10.1038/s41420-025-02551-4
Vorontsov I.E., Eliseeva I.A., Zinkevich A., Nikonov M., Abramov S., Boytsov A., Kamenets V., … Medvedeva Y.A., Jolma A., Kolpakov F., Makeev V.J., Kulakovskiy I.V. HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res*. 2024;52(D1):D154-D163. doi 10.1093/nar/gkad1077
Wingender E., Schoeps T., Dönitz J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res*. 2013;41(D1):D165-D170. doi 10.1093/nar/gks1123
Wingender E., Schoeps T., Haubrock M., Dönitz J. TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res*. 2015;43(D1):D97-D102. doi 10.1093/nar/gku1064
Wingender E., Schoeps T., Haubrock M., Krull M., Dönitz J. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res*. 2018;46(D1):D343-D347. doi 10.1093/nar/gkx987
Wu T., Hu E., Xu S., Chen M., Guo P., Dai Z., Feng T., Zhou L., Tang W., Zhan L., Fu X., Liu S., Bo X., Yu G. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation*. 2021;2(3):100141. doi 10.1016/j.xinn.2021.100141
Yang Y., Han X., Sun L., Shao F., Yin Y., Zhang W. ETS transcription factors in immune cells and immune-related diseases. *Int J Mol Sci*. 2024;25(18):10004. doi 10.3390/ijms251810004
Zambelli F., Pesole G., Pavesi G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief Bioinform*. 2013;14(2):225-237. doi 10.1093/bib/bbs016
Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nusbaum C., Myers R.M., Brown M., Li W., Liu X.S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137. doi 10.1186/gb-2008-9-9-r137

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила в редакцию 05.11.2025. После доработки 12.11.2025. Принята к публикации 17.11.2025.