doi 10.18699/bgrs2026-161

Обработка множественно картированных прочтений в данных РНК-ДНК взаимодействий

Звездин Д.[1, 2*], Жарикова А.[1, 2, 3], Миронов А.[1, 2]
[1] *Факультет биоинженерии и биоинформатики Московского государственного университета им. М.В. Ломоносова, Москва, Россия*
[2] *Институт общей генетики им. Н.И.Вавилова РАН, Москва, Россия*
[3] *Национальный медицинский исследовательский центр терапии и профилактической медицины Министерства здравоохранения Российской Федерации, Москва, Россия*
** zvezdindm@fbb.msu.ru*

Ключевые слова: хроматин; некодирующие РНК; РНК-ДНК взаимодействия; множественное картирование; EM-алгоритм

Мотивация и цель: Существенную часть транскриптов в эукариотических клетках представляют разнообразные классы некодирующих РНК. Функциональная роль большинства из них остается невыясненной, тогда как для некоторых установлена достаточно хорошо. За последние годы появился ряд экспериментальных протоколов, позволяющих фиксировать взаимодействия РНК с ДНК. Эти методы открыли новые возможности для изучения процессов, в которые вовлечены различные типы некодирующих РНК в ядрах эукариотических клеток. Все они основаны на технологиях секвенирования нового поколения. Для них характерна короткая длина прочтений, что приводит к тому, что существенная их часть картируется на геном неоднозначно и исключается из дальнейшего анализа. Потеря информации снижает покрытие данных и ограничивает анализ наблюдений, приходящих от повторов и дуплицированных генов. На данный момент не существует методов обработки таких множественно картированных чтений (МКЧ) для РНК-ДНК данных. Для данных Hi-C существует EM-алгоритм, который позволяет выбирать наиболее правдоподобную позицию картирования для таких чтений [1]. На его основании, мы разработали алгоритм, который позволяет сохранить существенную часть МКЧ для РНК-ДНК данных и повысить покрытие. Цель работы: применить разработанный алгоритм к различным данным РНК-ДНК взаимодействий и проанализировать полученные разрешенные позиции МКЧ.
Методы и алгоритмы: Для проверки гипотез в процессе разработки алгоритма мы выбрали набор данных RedChIP на клеточной линии человека K562 [2]. Для ускорения вычислений нами были отобраны чтения, для которых РНК и ДНК части попадают на одну из трех хромосом: 18, 19, 20. Данные были предобработаны с использованием конвейера обработки РНК-ДНК данных nf-rnachrom, реализованного на NextFlow. За основу мы выбрали подход, реализованный в алгоритме mHi-C, который решает поставленную задачу для данных ДНК-ДНК взаимодействий. Часть исходного алгоритма реализована с использованием базовых структур языка Python3, часть написана на языке Cython – компилируемом языке с похожем на Python синтаксисом. Для повышения эффективности вычислений и единообразия, мы переписали код с использованием высокопроизодительных Python библиотек: numpy и pandas. Также мы адаптировали алгоритм для работы с несимметричными контактными картами РНК-ДНК и заменили константную вероятность межхромосомных контактов на зависящую от пары хромосом. Мы добавили учет разной интенсивности, с которой РНК взаимодействуют с хроматином, в силу отличающегося уровня экспрессии.
Для оценки качества реализованного алгоритма мы предложили процедуру симуляции МКЧ с известной верной позицией. Процедура основана на генерации множественных картировок из части исходно уникальных путем добавления случайно выбранных позиций на геноме. Выбор числа картирований и их позиций у каждого такого симулированного чтения происходит на основании статистик, рассчитанных по наблюдаемым множественно картированным контактам.
Результаты: Мы провели анализ биологической значимости МКЧ. Мы проанализировали РНК-ДНК взаимодействия, чьи РНК части картируются уникально, и разделили их на полностью уникальные и с множественными позициями ДНК частей (мульти-ДНК). Мы разделили гены по биотипам, и для каждой РНК сравнили нормированные значения уникальных и мульти-ДНК контактов. Мы обнаружили 2 биотипа РНК, для которых характерна высокая представленность мульти-ДНК: малые ядрышковые РНК (мякРНК) и CDBox. мякРНК перепредставлены значимо (p-значение точного теста Фишера $5.23*10^{-13}$), число CDBox в выборке недостаточно для статистической оценки (5 РНК), но перепредставленность мульти-ДНК характерна для всех их представителей.
Для оценки качества разработанного нами алгоритма мы предложили процедуру симуляций МКЧ с известной верной позицией. С помощью симулированных данных мы оценили качество разрешения МКЧ по разным группам данных. В целом достигнуто хорошее качество – F1-мера равна 0.708, мульти-ДНК разрешаются хуже, чем мульти-РНК в силу разницы среднего числа позиций картировок на чтение (8.29 против 3.12). При этом наивная стратегия выбора случайной позиции для каждого чтения дает существенно худшее качество: F1-мера равна 0.267, что говорит о высокой информативности используемых в алгоритме подходов.
Мы применили алгоритм для разрешения наблюдаемых в эксперименте множественно картированных прочтений. Из выборки 132010 МКЧ (всего 759514 позиций) алгоритмом было выбрано 69447 позиций. Таким

образом, без проведения каких либо дополнительных экспериментов удалось получить прирост покрытия данных на 17.8 % (число уникальных контактов в выборке 390438). Выбранные алгоритмом позиции контактов по своим характеристикам, таким как зависимость интенсивности взаимодействий от линейного расстояния до гена РНК, схожи с исходно уникальными взаимодействиями. Построенные на их основании контактные РНК-ДНК карты не содержат артефактов, наблюдаемых на картах всех наблюдаемых позиций МКЧ (рисунок 1).

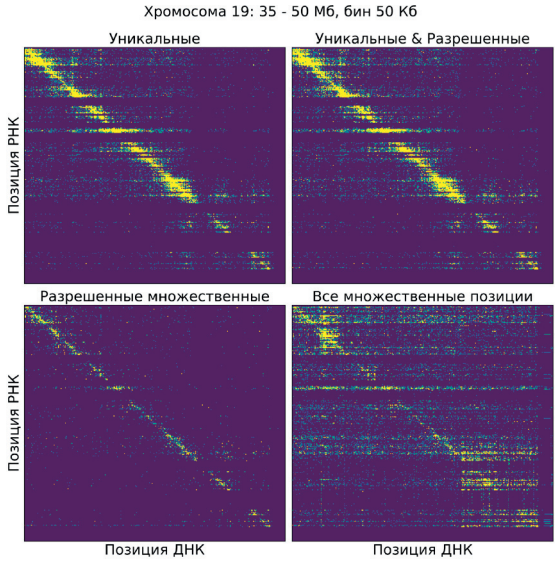


Рисунок 1. Контактные РНК-ДНК карты выборки реальных данных RedChIP K562.

Выводы: Мы показали, что есть биотипы РНК, склонные взаимодействовать с регионами повторов, однако стандартные методы исследования взаимодействий рассматривают лишь уникальные картирования. Это не позволяет изучать биологию подобных РНК, чья роль в хроматине связана с повторными регионами. Мы разработали алгоритм, решающий задачу поиска наиболее правдоподобных позиций картирования. Мы протестировали его работу на симулированных данных, и показали, что он определяет верные позиции с достаточным качеством. Применение алгоритма к наблюдаемым в эксперименте МКЧ, позволяет существенно повысить покрытие данных. Эти результаты говорят о полезности применения нашего подхода для анализа функциональной роли РНК-ДНК взаимодействий, обогащенных в повторных регионах.
Финансирование: Работа была поддержана FFRW-2024-0004 (125011400261-3).

Handling multi-mapped reads in RNA-DNA interaction data

Zvezdin D.[1, 2]*, Zharikova A.[1, 2, 3], Mironov A.[1, 2]
[1] *Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia*
[2] *Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia*
[3] *National Medical Research Center for Therapy and Preventive Medicine, Ministry of Healthcare of the Russian Federation, Moscow, Russia*
* *zvezdindm@fbb.msu.ru*

Key words: chromatin; non-coding RNA; RNA-DNA interactions; multiple mapping; EM algorithm

Motivation and Aim: A significant portion of transcripts in eukaryotic cells are represented by various classes of noncoding RNAs. The functional role of most of these remains unclear, while for some, it is fairly well established. In recent years, a number of experimental protocols have emerged that allow researchers to map RNA-DNA interactions. These methods have opened up new possibilities for studying the processes involving various types of noncoding RNAs in the nuclei of eukaryotic cells. All of these methods are based on next-generation sequencing technologies. They are characterized by short read lengths, which results in a significant portion of reads being mapped ambiguously to the genome and excluded from further analysis. This loss of information reduces data coverage and limits the analysis of observations from repeats and duplicated genes. Currently, no methods exist for processing such multiple mapped reads (MMRs) for RNA-DNA data. For Hi-C data, an EM algorithm exists that allows selecting the most plausible mapping position for such reads [1]. Based on this, we developed a method that allows us to preserve a significant portion of the MMRs for RNA-DNA data and improve data coverage. The goal of this study was to apply the developed algorithm to RNA-DNA interaction data and analyze the resulting resolved MMRs positions.
Methods and Algorithms: To test hypotheses during algorithm development, we selected the RedChIP dataset on the human K562 cell line [2]. To speed up computations, we selected reads for which the RNA and DNA parts fall on one of three chromosomes: 18, 19, and 20. The data was preprocessed using the nf-rnachrom RNA-DNA data processing pipeline, implemented in NextFlow. We based our approach on the mHi-C algorithm, which solves the problem for DNA-DNA interaction data. It is partially implemented with basic Python3 structures and with Cython, a compiled

language with a syntax similar to Python. To improve computational efficiency and consistency, we rewrote the code using high-performance Python libraries: numpy and pandas. We also adapted the algorithm to work with asymmetric RNA-DNA contact maps and replaced the constant probability of interchromosomal contacts with one that depends on the chromosome pair. We took into account different intensities with which RNAs interact with chromatin due to varying expression levels.

To assess the quality of the implemented algorithm, we proposed a procedure for simulating the MMRs with a known correct position. The procedure is based on generating multiple mappings from a portion of the initially unique ones by adding randomly selected positions on the genome. The number of mappings and their positions for each simulated read are selected based on statistics calculated from the observed multiple mapping contacts.

Results: We analyzed the biological significance of the MMRs. We analyzed RNA-DNA interactions whose RNA parts map uniquely and divided them into those with completely unique positions and those with multiple DNA positions (multi-DNA). We divided genes by biotype, and for each RNA, we compared the normalized values of unique and multi-DNA contacts. We found two RNA types characterized by a high representation of multi-DNA: small nucleolar RNAs (snoRNAs) and CDBox RNA. snoRNAs are significantly overrepresented (Fisher's exact test p-value of $5.23*10^{-13}$). The number of CDBoxes in the sample is insufficient for statistical evaluation (5 RNAs), but multi-DNA overrepresentation is a characteristic of all of them.

To evaluate the quality of our algorithm, we proposed a procedure for simulating the MMRs with a known correct position. Using the simulated data, we assessed the quality of the MMRs resolution. Overall, good quality was achieved: the F1-score is 0.708. Multi-DNA reads are resolved worse than multi-RNA reads due to the difference in the average number of mapping positions per read (8.29 vs. 3.12). A naive strategy for selecting a random position for each read yields significantly worse quality: the F1-score is 0.267, indicating the high information content of the approaches used in the algorithm.

We applied our method to resolve the multiple mapping reads observed in the experiment. From a sample of 132010 MMRs (759514 positions in total), the algorithm selected 69447 positions. Thus, without any additional experiments, we achieved a 17.8 % increase in data coverage (the number of unique contacts in the sample was 390438). The contact positions selected by the algorithm are similar to the originally unique interactions in their characteristics, such as the dependence of interaction intensity on the linear distance to the RNA gene. The constructed RNA-DNA contact maps do not contain the artifacts observed on the maps of all observed positions of the MMRs (Figure 1).

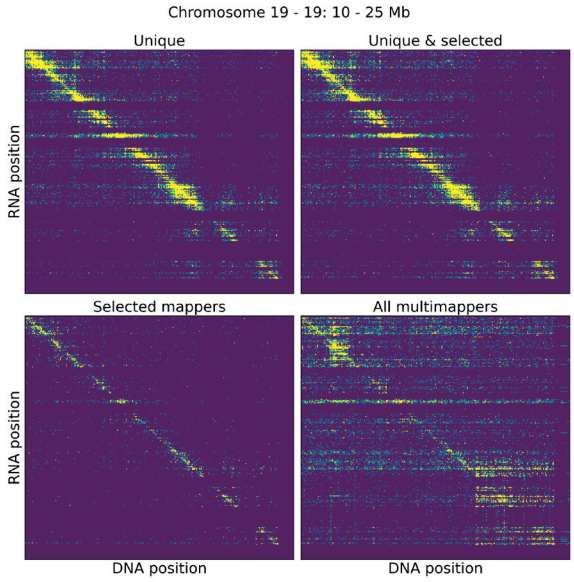


Figure 1. RNA-DNA contact maps of the RedChIP K562 data sample.

Conclusion: We demonstrated that there are RNA biotypes prone to interacting with repeat regions, but standard methods for studying interactions consider only unique mappings. This restricts studying the biology of the RNAs, whose role in chromatin is associated with repeat regions. We developed an algorithm for finding the most plausible mapping positions. We tested it on simulated data and demonstrated that it identifies the correct positions with sufficient quality. Applying the algorithm to the MMRs observed in the experiment significantly improves the data coverage. These results demonstrate the utility of our approach for analyzing the functional role of RNA-DNA interactions enriched in repeat regions.

Funding: This work was supported by FFRW-2024-0004 (125011400261-3).

Список литературы/References

1. Zheng, Y.; Ay, F.; Keles, S. Generative Modeling of Multi-Mapping Reads with mHi-C Advances Analysis of Hi-C Studies. eLife 2019, 8, e38070, doi:10.7554/eLife.38070.
2. Gavrilov, A.A.; Sultanov, R.I.; Magnitov, M.D.; Galitsyna, A.A.; Dashinimaev, E.B.; Aiden, E.L.; Razin, S.V. Brief Report: RedChIP Identifies Noncoding RNAs Associated with Genomic Sites Occupied by Polycomb and CTCF Proteins. Proc. Natl. Acad. Sci. U. S. A. 2022, 119, doi:10.1073/pnas.2116222119.