Биомедицинская химия, 2017 том 63, вып. 5, с. 418-422.

КОМПЬЮТЕРНЫЕ МЕТОДЫ АНАЛИЗА ХРОМОСОМНЫХ КОНТАКТОВ В ЯДРЕ КЛЕТКИ ПО ДАННЫМ ТЕХНОЛОГИЙ СЕКВЕНИРОВАНИЯ

Ю.Л. Орлов[1,2]*, О. Тьерри[1,3], А.Г. Богомолов[1,4], А.В. Цуканов[1], Е.В. Кулакова[1], Э.Р. Галиева[1], А.О. Брагин[4], Г. Ли[5]

[1]Новосибирский государственный университет,
Новосибирск, 630090, ул. Пирогова, 2,; эл. почта: orlov@bionet.nsc.ru
[2]Институт морских биологических исследований, Севастополь
[3]Университет Бордо, Бордо, Франция
[4]Институт цитологии и генетики СО РАН, Новосибирск
[5]Аграрный Университет Хуажонг, Ухань, Китай

Организация пространственной структуры хромосом, их укладки в интерфазном ядре клетки, установление взаимодействующих участков генома эукариот, физически контактирующих друг с другом, активно изучается в мире с использованием современных технологий секвенирования. Основным методом исследования укладки хромосом с помощью секвенирования пар контактирующих фрагментов ДНК стал Hi-C (метод определения конформации хромосом высокого порядка). Исследование взаимодействий хроматина, трехмерной структуры генома и её воздействия на регуляцию транскрипции позволяет понять фундаментальные биологические процессы с точки зрения структурной регуляции экспрессии генов в клетке, что важно для изучения рака. С помощью методов, основанных на иммунопреципитации хроматина и последующем секвенировании (ChIP-seq), стало возможным определение сайтов связывания транскрипционных факторов, регулирующих транскрипцию генов в геномах эукариот; созданы геномные карты связывания транскрипционных факторов. Технология ChIA-PET (Chromatin ImmunoPrecipitation Analysis - Pair End Tags) позволяет исследовать не только отдельные сайты связывания, но и пары таких сайтов на хромосомах, расположенные рядом в трёхмерном пространстве ядра клетки, и находящиеся в комплексе с исследуемым белком. В данной работе рассмотрены принципы построения геномных карт и матриц хромосомных контактов по данным технологий ChIA-PET и Hi-C. Представлен обзор существующих программных решений и компьютерных инструментов для анализа трёхмерной структуры генома по данным Hi-C и ChIA-PET.

Ключевые слова: хромосомные контакты, программное обеспечение, секвенирование, регуляция транскрипции, Hi-C, ChIA-PET

DOI: 10.18097/PBMC20176305418

ВВЕДЕНИЕ

Исследование хромосомных контактов в интерфазном ядре клетки имеет огромное значение для понимания молекулярных механизмов регуляции экспрессии генов в масштабе генома, хромосомном регуляторном коде и его влиянии на развитие рака [1]. Развитие новых экспериментальных технологий, основанных на высокопроизводительном секвенировании ДНК, даёт возможность картирования хромосомных контактов [2]. Методы исследования хромосомных контактов в геноме – Hi-C [3] и ChIA-PET [4] – активно развиваются с 2009 года. Технология Hi-C (High conformation Capture – устоявшийся термин от английского "конформации хромосом высокого порядка") [3] позволяет определять все хромосомные контакты в ядре клетки. Технология ChIA-PET (Chromatin Interaction Analysis by Paired-End-Tag sequencing – анализ взаимодействий хроматина с помощью секвенирования парных концов) включает стадию иммунопреципитации хроматина и предназначена для исследования хромосомных контактов, опосредованных тем или иным белком [4]. Данная работа представляет обзор компьютерных методов обработки данных хромосомных контактов [5, 6].

В последние годы с использованием технологий Hi-C и родственных методов получены новые знания об особенностях трёхмерной архитектуры (укладки) генома человека в интерфазном ядре клетки, влияющих на регуляцию экспрессии генов [7, 8, 9]. Рассматриваются вопросы формирования топологических доменов, выделения петель, формируемых контактами когезина и фактора CTCF [10]. Для обработки большого объёма данных ChIA-PET и Hi-C, развиваются компьютерные инструменты анализа, позволяющие получить качественно новую информацию о различных аспектах структурной организации генома [11].

1. КАРТЫ ХРОМОСОМНЫХ КОНТАКТОВ И СРАВНЕНИЕ ТЕХНОЛОГИЙ

Трехмерные координаты протяженной последовательности ДНК могут быть представлены двухмерной картой контактов участков этой последовательности для последующего моделирования, так же как и полипептидная цепочка может быть представлена матрицей контактов аминокислотных остатков. В свою очередь, двухмерная карта (симметричная матрица) контактов участков (звеньев) последовательности позволяет реконструировать трёхмерную укладку [11]. Основной технической

* - адресат для переписки

задачей определения структуры хромосомы является построение матрицы контактов участков хромосомы. Секвенирование парных контактов выполняется на основе анализа ДНК из популяции клеток, поэтому выводы о структуре контактов в геноме носят статистический характер.

Схематический пример построения такой карты показан на рисунке. Матрица симметричная, аналогична дот-матрице при выравнивании последовательностей. В такой матрице 1 соответствует контакту в данной позиции последовательности, 0 – отсутствию контакта. Размер шага (участка) последовательности хромосомы при построении матрицы составляет десятки тысяч нуклеотидов. Получающиеся матрицы контактов затем анализируются с помощью компьютерных программ, проводятся вычислительные преобразования, кластеризация участков хромосомы, выделяются топологические ассоциированные домены, районы генома [12].

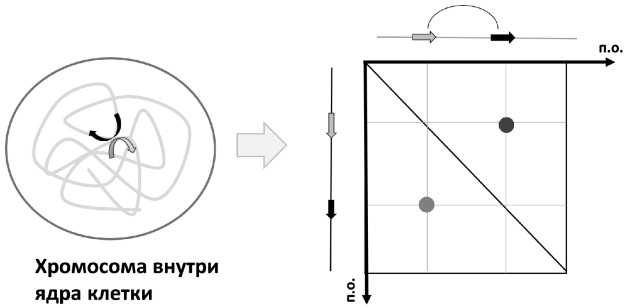

Рисунок. Схема представления контактов в хромосоме: от взаимодействующих участков в ядре клетки (слева) к построению матрицы контактов (справа). Стрелками помечены контактирующие участки хромосомы в трёхмерном пространстве и их представление в матрице.

Построение, обработка и анализ таких карт контактов определяют компьютерные задачи и программные решения – процессинг данных, хранение и визуализация матриц контактов большого размера, выделение топологических доменов [13].

Метод ChIA-PET более специфичен для определения контактирующих участков хромосом, контакты которых опосредованы белками или молекулярными комплексами, например комплексом РНК-полимеразы II [7]. ChIA-PET является развитием технологий ChIP-on-chip, ChIP-PET и ChIP-seq для определения ДНК-белкового связывания в геноме [4, 7]. Кроме того, ChIA-PET позволяет определить петли хромосом в геноме, определяя отдельный класс компьютерных задач моделирования таких петель [14].

Технология Hi-C позволяет определить все контактирующие участки в геноме, безотносительно к какому-либо белку, задавая направление компьютерного моделирования глобальной укладки хромосом в интерфазном ядре клетки [3]. Разработаны технологические продолжения Hi-C – метод TCC (Tethered Conformation Capture) [15], метод Capture Hi-C [16]. Районы активно транскрибируемых генов чаще участвуют в межхромосомных контактах. Этот факт объясняют выделением петель хромосом в область концентрации белков транскрипционного аппарата (так называемые “фабрики транскрипции”), где становится возможным контакт с другим активным районом [13, 17]. Модель “фрактальной глобулы” позволяет объяснить компактизацию длинной молекулы ДНК в ядре клетки [3]. Дальнейшие исследования показали отклонения от модели идеальной фрактальной глобулы и наличие топологических доменов на хромосоме [18]. Как альтернативу методам секвенирования отметим новые подходы микроскопии сверхвысокого разрешения для исследования структуры хромосом в ядре – 3D STORM (stochastic optical reconstruction microscopy) [19].

Изучение трёхмерной организации генома важно для анализа геномной регуляции на клеточном уровне при многих заболеваниях [20], помогает исследовать рак [21, 22].

2. БАЗЫ ДАННЫХ ТРЁХМЕРНЫХ СТРУКТУР ХРОМОСОМ

При обработке данных Hi-C выделяются пространственные домены на хромосоме: такая информация представлена в Интернет-доступных базах данных, таких как 3DGD [23], 4DGenome [24], 3CDB [25].

База данных 3DGD (Three dimensional genome database) обеспечивает доступ и визуализацию трехмерной структуры хроматина Hi-C.

База данных 4DGenome [24] содержит данные взаимодействий хроматина, собранные из литературных источников, включая данные методов 3C, 4C-Seq, 5C, Hi-C, ChIA-PET и Capture-C. В 4DGenome содержится около 8 миллионов записей для более 100 типов клеток и тканей. Отметим курируемую базу данных хромосомных контактов 3CDB [25], содержащую информацию о взаимодействиях хроматина для 17 видов из более 5000 научных публикаций.

База данных CTCFBSDB 2.0 (http://insulatordb.uthsc.edu/) содержит информацию о расположении хромосомных доменов, ограниченных инсулятором CTCF (CCCTC-binding factor – транскрипционный фактор, определяющий границы транскрипции на хромосоме) [26].

3. КОМПЬЮТЕРНЫЕ ИНСТРУМЕНТЫ АНАЛИЗА ТРЁХМЕРНОЙ ОРГАНИЗАЦИИ ГЕНОМА ПО ДАННЫМ СЕКВЕНИРОВАНИЯ

В связи с задачами обработки данных о трёхмерной организации генома возник ряд компьютерных методов анализа данных Hi-C и ChIA-PET; одна из наиболее полных версий списка компьютерных инструментов регулярно обновляется на сайте OMICtools (http://omictools.com/3c-4c-5c-hi-c-chia-pet-category). Таблица содержит основные инструменты, которые можно разделить на средства анализа данных ChIA-PET, Hi-C, геномные браузеры и средства для анализа 3C-методов.

Рассмотрим более подробно существующие в мире программные инструменты для анализа данных ChIA-PET (таблица).

ChIA-PET Tool (http://chiapet.gis.a-star.edu.sg/downloads/chia-pet-tools) – исторически первый пакет программного обеспечения для обработки данных ChIA-PET [27]. Пакет предназначен для операционных систем Unix/Linux, имеет открытый исходный код. Недавно разработан новый инструмент ChIA-PET2, предназначенный для конвейерной обработки различных форматов ДНК ChIA-PET [5].

Пакет Mango (https://github.com/dphansti/mango) оценивает статистическую достоверность оценок хромосомных контактов [28], для анализа экспериментов ChIA-PET.

Основные инструменты для работы с данными Hi-C [3] включают пакет HiClib (https://bitbucket.org/mirnylab/hiclib) для построения матриц контактов, их нормализации [32].

Программа HiBrowse (https://hyperbrowser.uio.no/3d/), реализованная на языке Python, представляет собой веб-сервер, выполняющий статистический анализ, интерпретацию и визуализацию данных, полученных методами Hi-C, TCC и др. [30].

Программы Juicer и Juicebox представляют компьютерные решения для визуализации контактов по данным Hi-C [33] (Rowley 2016).

Пакет CHiCAGO (http://regulatorygenomicsgroup.org/chicago) определяет петли хроматина данные по данным технологии CHi-C (Capture Hi-C) [16].

ЗАКЛЮЧЕНИЕ

Вычислительная сложность анализа данных о регуляторных районах генов увеличивается при рассмотрении взаимодействий между генами в пространстве ядра клетки, что требует разработки новых программных средств и адаптации программных конвейеров [12].

Исследование хромосомных контактов не только подтверждает иерархичность организации топологических доменов в геноме человека и близость регуляторных последовательностей к транскрибируемым генам в пространстве ядра, но и позволяет по-новому изучать регуляцию экспрессии генов, через общие категории генных онтологий, паттерны совместной экспрессии. Продолжаются работы по исследованию хромосомной упаковки, определению дистальных взаимодействий хроматина в модельных организмах [34].

Дальнейшие исследования пространственной архитектуры с помощью новых данных ChIA-PET в клеточных линиях, улучшат понимание транскрипционной регуляции при развитии заболеваний [21, 22, 35]. Применения методов

Таблица. Программы анализа данных секвенирования для исследования трёхмерной структуры генома

Инструмент	Описание	Интернет-ссылки. Публикации
ChIA-PET		
ChIA-PET Tool	Пакет программного обеспечения для автоматической обработки данных о последовательностях, полученных методом ChIA-PET.	http://chiapet.gis.a-star.edu.sg/ [27]
Mango	Утилита для данных, полученных методом ChIA-PET. Рассчитывает статистическую достоверность оценок взаимодействия.	https://github.com/dphansti/mango [28]
ChIA-PET2	Конвейер для анализа данных ChIA-PET, включая данные, сгенерированные по различным экспериментальным протоколам	https://github.com/GuipengLi/ChIA-PET2. [5]
Hi-C		
FisHiCal	Интегрирует данные Hi-C и FISH, инструмент для анализа пространственной структуры хромосом	http://cran.r-project.org/web/packages/FisHiCal/index.html [29]
HiBrowse	Web Toolkit для анализа, интерпретации и визуализации данных Hi-C, TCC, GCC и др.	https://hyperbrowser.uio.no/3d/ [30]
HIPPIE	Конвейер с высокой пропускной способностью для обнаружения промоторов,взаимодействующих с энхансерами.	http://wanglab.pcbi.upenn.edu/hippie/ [31]
Hiclib ICE	Вычислительная утилита объединяет средства картирования прочтений ДНК и метод управления данными с итерационной коррекцией. Полногеномные карты основаны на вероятности контактов.	http://mirnylab.bitbucket.org/hiclib/index.html [32]
Juicer и Juicebox	Компьютерные решения для визуализации контактов по данным Hi-C.	http://aidenlab.org/juicer/ [33]

анализа трёхмерной структуры генома открывают новые области в биотехнологии, требуя также развития специализированных программ [34].

БЛАГОДАРНОСТИ

Авторы благодарны Н.Р. Баттулину, А.И. Дергилеву, Н.Л. Подколодному за предоставление данных, вычислительных ресурсов и техническую поддержку. Работа поддержана РФФИ и бюджетным проектом ИЦиГ СО РАН 0324-2016- 0008.

ЛИТЕРАТУРА

1. *Игнатьева Е.В., Подколодная О.А., Орлов Ю.Л., Васильев Г.В., Колчанов Н.А.* (2015) Генетика, **51**(4), 409-429.
2. *Dekker J., Rippe K., Dekker M., Kleckner N.* (2002) Science, **295**(5558), 1306-1311.
3. *Lieberman-Aiden E., Van Berkum N.L., Williams L., Imakaev M., Ragoczy T., Telling A., Amit I., Lajoie B.R., Sabo P.J., Dorschner M.O.* (2009) Science, **326**(5950), 289-293.
4. *Fullwood M.J., Liu M.H., Pan Y.F., Liu J., Xu H., Mohamed Y.B., Orlov Y.L., Velkov S., Ho A., Mei P.H. et al.* (2009) Nature, **462**(7269), 58-64.
5. *Li G., Chen Y., Snyder M.P., Zhang M.Q.* (2017) Nucl. Acids Res., **45**(1), e4. DOI:10.1093/nar/gkw809
6. *Szalaj P., Tang Z., Michalski P., Pietal M.J., Luo O.J., Sadowski M., Li X., Radew K., Ruan Y., Plewczynski D.* (2016) Genome Res., **26**(12), 1697-1709.
7. *Li G., Ruan X., Auerbach R.K., Sandhu K.S., Zheng M., Wang P., Poh H.M., Goh Y., Lim J., Zhang J. et al.* (2012) Cell, **148**(1-2), 84-98.
8. *Li G., Cai L., Chang H., Hong P., Zhou Q., Kulakova E.V., Kolchanov N.A., Ruan Y.* (2014) BMC genomics, **15**(Suppl 12), S11. DOI: 10.1186/1471-2164-15-S12-S11.\
9. *Ramani V., Cusanovich D.A., Hause R.J., Ma W., Qiu R., Deng X., Blau C.A., Disteche C.M., Noble W.S., Shendure J., Duan Z.* (2016) Nat Protoc., **11**(11), 2104-2121.
10. *Tang Z., Luo O.J., Li X., Zheng M., Zhu J.J., Szalaj P., Trzaskoma P., Magalska A., Wlodarczyk J., Ruszczycki B. et al.* (2015) Cell, **163**(7), 1611-1627.
11. *Chiariello A.M., Annunziatella C., Bianco S., Esposito A., Nicodemi M.* (2016) Sci. Rep., **6**, 29775.
12. *Кулакова Е.В., Спицина А.М., Орлова Н.Г., Дергилев А.И., Свичкарев А.В., Сафронова Н.С., Черных И.Г., Орлов Ю.Л.* (2015) Программные системы: теория и приложения, **6**(2), 129-148.
13. *Ay F., Noble W.S.* (2015) Genome Biol., **16**(1), 1-15.
14. *Ulianov S.V., Khrameeva E.E., Gavrilov A.A., Flyamer I.M., Kos P., Mikhaleva E.A., Penin A.A., Logacheva M.D., Imakaev M.V., Chertovich A., Gelfand M.S., Shevelyov Y.Y., Razin S.V.* (2016) Genome Res., **26**(1), 70-84.
15. *Sanyal A., Bau D., Marti-Renom M.A., Dekker J.* (2011) Curr. Opin. Cell Biol., **23**(3), 325-331.
16. *Cairns J., Freire-Pritchett P., Wingett S.W., Varnai C., Dimond A., Plagnol V., Zerbino D., Schoenfelder S., Javierre B.M., Osborne C., Fraser P., Spivakov M.* (2016) Genome Biol., **17**(1), 127.
17. *Battulin N., Fishman V.S., Mazur A.M., Pomaznoy M., Khabarova A.A., Afonnikov D.A., Prokhortchouk E.B., Serov O.L.* (2015) Genome Biol., **16**, 77.
18. *Wang S., Su J.H., Beliveau B.J., Bintu B., Moffitt J.R., Wu C.T., Zhuang X.* (2016) Science, **353**(6299), 598-602.
19. *Georgieva M., Cattoni D.I., Fiche J.B., Mutin T., Chamousset D., Nollmann M.* (2016) Methods, **105**, 44-55.
20. *Grubert F., Zaugg J.B., Kasowski M., Ursu O., Spacek D.V., Martin A.R., Greenside P., Srivas R., Phanstiel D.H., Pekowska A.* (2015) Cell, **162**(5), 1051-1065.
21. Babu D., Fullwood M.J. (2015) Nucleus, **6**(5), 382-393.
22. *Michailidou K., Beesley J., Lindstrom S., Canisius S., Dennis J., Lush M.J., Maranian M.J., Bolla M.K., Wang Q., Shah M.* (2015) Nat Genet., **47**(4), 373-380.
23. *Li C., Dong X., Fan H., Wang C., Ding G., Li Y.* (2014) Bioinformatics, **30**(11), 1640-1642.
24. *Teng L., He B., Wang J., Tan K.* (2015) Bioinformatics, **31**(15), 2560-2564.
25. *Yun X., Xia L., Tang B., Zhang H., Li F., Zhang Z.* (2016) Database (Oxford). pii: baw044. DOI:10.1093/database/baw044
26. *Ziebarth J.D., Bhattacharya A., Cui Y.* (2013) Nucl. Acids Res., **41**(D1), D188-D194.
27. *Li G., Fullwood M., Xu H., Mulawadi F., Velkov S., Vega V., Ariyaratne P., Mohamed Y., Ooi H., Tennakoon C.* (2010) Genome Biol., **11**, R22. DOI:10.1186/gb-2010-11-2-r22
28. *Phanstiel D.H., Boyle A.P., Heidari N., Snyder M.P.* (2015) Bioinformatics, **31**(19), 3092-3098.
29. *Shavit Y., Hamey F.K., Lio P.* (2014) Bioinformatics, **30**(21), 3120-3122.
30. *Paulsen J., Sandve G.K., Gundersen S., Lien T.G., Trengereid K., Hovig E.* (2014) Bioinformatics, **30**(11), 1620-1622.
31. *Hwang Y.C., Lin C.F., Valladares O., Malamon J., Kuksa P.P., Zheng Q., Gregory B.D., Wang L.S.* (2015) Bioinformatics, **31**(8), 1290-1292.
32. *Imakaev M., Fudenberg G., McCord R.P., Naumova N., Goloborodko A., Lajoie B.R., Dekker J., Mirny L.A.* (2012) Nat. Methods, **9**(10), 999-1003.
33. *Rowley M.J, Corces V.G.* (2016) Mol. Cell, **64**(1), 9-11.
34. *Zolotarev N., Fedotova A., Kyrchanova O., Bonchuk A., Penin A.A., Lando A.S., Eliseeva I.A., Kulakovskiy I.V., Maksimenko O., Georgiev P.* (2016) Nucl. Acids Res., **44**(15), 7228-7241.
35. *Li R., Liu Y., Li T., Li C.* (2016) Sci Rep., **6**, 34651.

Поступила: 31. 08. 2017.
Принята к печати: 14. 09. 2017.

COMPUTER METHODS OF ANALYSIS OF CHROMOSOME CONTACTS IN THE CELL NUCLEUS BASED ON SEQUENCING TECHNOLOGY DATA

Y.L. Orlov[1,2], O. Thierry[1,3], A.G. Bogomolov[1,4], A.V. Tsukanov[1], E.V. Kulakova[1], E.R. Galieva[1], A.O. Bragin[4], G. Li[5]

[1]Novosibirsk State University, Novosibirsk, Russia; e-mail: orlov@bionet.nsc.ru
[2]Marine Biology Research Institute, Sevastopol, Russia
[3]University of Bordeaux, Bordeaux, France
[4]Institute of Cytology and Genetics, Novosibirsk, Russia
[5]Huazhong Agricultural University, Wuhan, Hubei, China

The study spatial chromosome structure and chromosome folding in the interphase cell nucleus is an important challenge of world science. Detection of eukaryotic genome regions that physically interact with each other could be done by modern sequencing technologies. A basic method of chromosome folding by total sequencing of contacting DNA fragments is HI-C. Long-range chromosomal interactions play an important role in gene transcription and regulation. The study of chromosome interactions, 3D (three-dimensional) genome structure and its effect on gene transcription allows revealing fundamental biological processes from a viewpoint of structural regulation and are important for cancer research. The technique of chromatin immunoprecipitation and subsequent sequencing (ChIP-seq) make possible to determine binding sites of transcription factors that regulate expression of eukaryotic genes; genome transcription factors binding maps have been. The ChIA-PET technology allows exploring not only target protein binding sites, but also pairs of such sites on proximally located and interacting with each other chromosomes co-located in three-dimensional space of the cell nucleus. Here we discuss the principles of the construction of genomic maps and matrices of chromosome contacts according to ChIA-PET and Hi-C data that capture the chromosome conformation and overview existing software for 3D genome analysis including in house programs of gene location analysis in topological domains.

Key words: chromosomal contacts, software, sequencing, transcription regulation, Hi-C, ChIA-PET