Тема работы:
Анализ пространственной организации хроматина и эпигенетических профилей ооцитов птиц методом single-cell секвенирования.

Состав коллектива:
Лагунов Тимофей Аркадьевич, аспирант ФФ НГУ, инженер-исследователь ИЦиГ СО РАН
Фишман Вениамин Семёнович, д.б.н., в.н.с. ИЦиГ СО РАН, доцент НГУ

Информация о гранте:
РНФ: Разработка новых подходов для исследования механизмов пространственной организации хроматина и их функционального значения в регуляции генной экспрессии у животных (РНФ 22-14-00247), 2022-2024, рук. Фишман В.С.

Научное содержание работы:

1) Постановка задачи:
Работа направлена на исследование хроматиновой архитектуры и эпигенетических особенностей ооцитов курицы (Gallus gallus) на стадии ламповых щеток. Ставились следующие задачи:

- Подготовка и анализ данных single-cell Hi-C
- сравнение протоколов single-cell Hi-C: PCR-основанный и MDA-основанный;
- Подготовка и анализ данных Methyl-seq;
- анализ распределения метилирования в геноме курицы;
- разработка и применение HMM-модели для сегментации профиля метилирования;
- Сопоставление профиля метилирования между ооцитами и другими клетками курицы

2) Современное состояние проблемы:
Наиболее полно процесс мейоза описан у млекопитающих из-за большей изученности этого таксона. Половые клетки других модельных организмов обладают уникальными особенностями, изучение которых важно, как для понимания фундаментальных механизмов функционирования и эволюции мейотических хромосом, так и для решения прикладных задач в области генетики животных. На стадии диплотены мейоза геном птиц представлен уникальной структурой хромосом типа ламповых щеток (ЛЩ). Несмотря на то, что сами хромосомы типа ламповых щеток были описаны более ста лет назад, механизм, лежащий в основе их формирования и функциональное значение этих структур, до сих пор остается загадкой, на разрешение которой направлен наш проект.
Для получения информации о пространственной конформации хромосом типа ЛЩ с большим геномным разрешением, необходимо использовать технологию Hi-C. Поскольку ооциты на стадии ламповых щёток курицы – объект, который можно получить из животного лишь в штучном количестве, то необходимо использовать модифицированный протокол для работы с одиночными клетками (single cell Hi-C, scHi-C). Протоколы scHi-C можно разделить на две основные категории по способу амплификации химерных последовательностей ДНК после лигирования: 1) амплификация методом PCR после фрагментации, пришивания адаптеров и обогащения (PCR-основанный метод, как в изначальном Hi-C; [1–4]); 2) метод множественной параллельной амплификации со случайных гексамерных праймеров полимеразой с высокой процессивностью (англ. multiple displacement amplification, MDA, [5; 6]). Для дальнейшей работы с анализом пространственной конформации необходимо сопоставить методы и выбрать оптимальный для поставленной задачи.
Одной из особенностей хромосом типа ламповых щёток является гипертраснкрипция. В ооцитах млекопитающих метилирование зависит от транскрипционной активности,

охватывая, как правило, тела активных генов, тогда как промоторы и интергенные области остаются гипометилированными ([7]). Для анализа профиля метилирования используют Methyl-seq анализ. Несмотря на высокую транскрипционную активность в диплотенных ооцитах птиц и формирование уникальной архитектуры хроматина в виде ламповых щёток, данные о распределении метилирования ДНК в женских половых клетках птиц до настоящего времени отсутствовали. При этом уже показано, что геном сперматозоидов курицы гипометилирован по сравнению с соматическими клетками, что связывают с отсутствием кофактора DNMT3L в геноме курицы ([8]).

1. Ke Y., Xu Y., Chen X., Feng S., Liu Z., Sun Y., Yao X., Li F., Zhu W., Gao L., Chen H., Du Z., Xie W., Xu X., Huang X., Liu J. 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis//Cell, 2017, T. 170, N 2, C. 367-381.e20.
2. Stevens T.J., Lando D., Basu S., Atkinson L.P., Cao Y., Lee S.F., Leeb M., Wohlfahrt K.J., Boucher W., O'Shaughnessy-Kirwan A., Cramard J., Faure A.J., Ralser M., Blanco E., Morey L., Sansó M., Palayret M.G.S., Lehner B., Di Croce L., Wutz A., Hendrich B., Klenerman D., Laue E.D. 3D structures of individual mammalian genomes studied by single-cell Hi-C//Nature, 2017, Vol. 544, No. 7648, P. 59-64.
3. Collombet S., Ranisavljevic N., Nagano T., Varnai C., Shisode T., Leung W., Piolot T., Galupa R., Borensztein M., Servant N., Fraser P., Ancelin K., Heard E. Parental-to-embryo switch of chromosome organization in early embryogenesis//Nature, 2020, Vol. 580, No. 7801, P. 142-146.
4. Nagano T., Lubling Y., Várnai C., Dudley C., Leung W., Baran Y., Mendelson Cohen N., Wingett S., Fraser P., Tanay A. Cell-cycle dynamics of chromosomal organization at single-cell resolution//Nature, 2017, Vol. 547, No. 7661, P. 61-67.
5. Ulianov S.V., Zakharova V.V., Galitsyna A.A., Kos P.I., Polovnikov K.E., Flyamer I.M., Mikhaleva E.A., Khrameeva E.E., Germini D., Logacheva M.D., Gavrilov A.A., Gorsky A.S., Nechaev S.K., Gelfand M.S., Vassetzky Y.S., Chertovich A.V., Shevelyov Y.Y., Razin S.V. Order and stochasticity in the folding of individual Drosophila genomes//Nature Communications, 2021, Vol. 12, No. 1, P. 41.
6. Flyamer I.M., Gassler J., Imakaev M., Brandão H.B., Ulianov S.V., Abdennur N., Razin S.V., Mirny L.A., Tachibana-Konwalski K. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition//Nature, 2017, Vol. 544, No. 7648, P. 110-114.
7. Sendžikaitė G., Kelsey G. The role and mechanisms of DNA methylation in the oocyte//Essays in Biochemistry, 2019, Vol. 63, No. 6, P. 691-705.
8. Greenberg M.V.C., Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease//Nature Reviews Molecular Cell Biology, 2019, Vol. 20, No. 10, P. 590-607.

3) Подробное описание работы, включая используемые алгоритмы:

В работе использовались современные алгоритмы обработки и анализа данных.

Обработка данных Hi-C-эксперимента

Первичная обработка сырых данных секвенирования включала удаление адаптеров Illumina с использованием программы Cutadapt версии 4.1 [1]. Для выравнивания парных чтений использовался скрипт juicer версии 1.5.6 [2]. Мы использовали сборку генома курицы ASM2420605v2 [3]. При фильтрации оставлялись только пары чтений без дупликатов и с качеством картирования не ниже 30.

Статистика качества рассчитывалась по выходным данным juicer с незначительными модификациями в расчёте отношения между внутри- и межхромосомными контактами, как описано ранее [4]. Производился анализ количества уникальных контактов для каждого фрагмента, ограниченного сайтами рестрикции DpnII. Количество контактов не должно превышать 8 из соображений, что каждый фрагмент представлен 4 копиями (две

гомологичные хромосомы содержат по две сестринские хроматиды), и у каждого фрагмента 2 конца для лигирования с другим фрагментом. Для построения контактных матриц в формате mcool использовался вывод juicer в сочетании с пакетом cooler версии 0.10.2 [5]. В связи с относительно низким числом контактов на клетку (~500 тыс.), функция нормализации (balance) в cooler не применялась. Вместо этого, весам бинов с нулевым покрытием присваивалось значение “nan”, а весам бинов с покрытием — значение 1. Агрегированные графики строились с использованием пакета coolpuppy версии 1.1.0 [6] с параметром maxshift = 2 Мб и флагом “local”. Расчёт кривых зависимости вероятности контакта от геномного расстояния осуществлялся с помощью пакета cooltools [7] с параметром сглаживания Гаусса sigma = 0.1.

Программы для анализа количества артефактов были написаны собственноручно на языке python 3.9.

Анализ метилирования

Удаление адаптеров Illumina из сырых прочтений выполнялось с помощью cutadapt версии 4.1 [1]. Для обработки прочтений Methyl-seq-эксперимента использовались программы Bismark версии 0.23.0 [8; 9] и Bowtie2 версии 2.4.4 [10–13] с параметром -t, равным 10. Прочтения выравнивались на сборку ASM2420605v2 [3], при этом отбрасывались прочтения с качеством картирования < 10. Профили метилирования CpG рассчитывались на основе выравниваний с помощью bismark-methylation-extractor.

Для анализа профиля метилирования нами была модифицирована Python-реализация скрытой марковской модели (HMM), предложенная в работе [14]. Эмиссионные вероятности состояний модели описывались биномиальным распределением $Bin(n_t, b_k)$, где n_t— глубина покрытия по CpG-сайту t, а b_k — средний уровень метилирования, соответствующий состоянию k. Для повышения влияния позиций с высоким покрытием на обучение модели, биномиальное распределение возводилось в степень $n_t + 2$, как предложено в работе [15].

Также была модифицирована матрица переходов между состояниями с учётом снижения корреляции уровней метилирования при увеличении расстояния между соседними CpG-сайтами (аналогично подходу [16]):

$$P(S_t|S_{t-1} = k, d_{t-1}) = \begin{cases} \frac{1}{4} - \frac{1}{4}\exp\left(-\frac{d_{t-1}}{d_0}\right), & S_t \neq S_{t-1} \\ \frac{1}{4} + \frac{3}{4}\exp\left(-\frac{d_{t-1}}{d_0}\right), & S_t = S_{t-1} \end{cases}$$

где S_t — состояние CpG-сайта t, S_{t-1} — состояние CpG-сайта $t-1$, d_{t-1} — расстояние между сайтами $t-1$ и t, d_0 — параметр затухания корреляции, k – номер состояния.

В отличие от классических HMM, предназначенных для работы с направленными временными рядами, модель метилирования должна учитывать, как предшествующий, так и последующий сайты. Для реализации этого была добавлена обратная (реверсивная) выборка в процессе обучения и декодирования. После выполнения декодирования в прямом и обратном направлениях, были выделены позиции, в которых предсказанные состояния не совпадали. Эти позиции помечались как "неопределённые" и исключались из дальнейшего анализа.

Формирование обучающих и тестовых выборок выполнялось скриптом set_maker.py. Для калибровки параметров модели были отобраны 54 непересекающихся участка длиной по 300 тыс. пар оснований, случайно распределённых по геному таким образом, чтобы в обучающей и тестовой выборках были равномерно представлены макро-, промежуточные и микрохромосомы курицы. Для выбора оптимальных параметров модели использовались информационные критерии Акаике (AIC) и Байеса (BIC) из реализации HMM [14; 17]. Оптимальными значениями параметров оказались: число состояний n=3, параметр расстояния $d_0 = 1250$. Эти параметры были использованы для полного декодирования метилирования в данных по ооцитам.

1. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads//EMBnet.journal, 2011, T. 17, N 1, C. 10.
2. Durand N.C., Shamim M.S., Machol I., Rao S.S.P., Huntley M.H., Lander E.S., Aiden E.L. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments//Cell Systems, 2016, Vol. 3, No. 1, P. 95-98.
3. Huang Z., Xu Z., Bai H., Huang Y., Kang N., Ding X., Liu J., Luo H., Yang C., Chen W., Guo Q., Xue L., Zhang X., Xu L., Chen M., Fu H., Chen Y., Yue Z., Fukagawa T., Liu S., Chang G., Xu L. Evolutionary analysis of a complete chicken genome//Proceedings of the National Academy of Sciences, 2023, Vol. 120, No. 8, P. e2216641120.
4. Gridina M., Mozheiko E., Valeev E., Nazarenko L.P., Lopatkina M.E., Markova Z.G., Yablonskaya M.I., Voinova V.Y., Shilova N.V., Lebedev I.N., Fishman V. A cookbook for DNase Hi-C//Epigenetics & Chromatin, 2021, Vol. 14, No. 1, P. 15.
5. Abdennur N., Mirny L.A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays//Bioinformatics, 2020, Vol. 36, Cooler, No. 1, P. 311-316.
6. Flyamer I.M., Illingworth R.S., Bickmore W.A. Coolpup.py: versatile pile-up analysis of Hi-C data//Bioinformatics, 2020, Vol. 36, <i>Coolpup.py, No. 10, P. 2980-2985.
7. Open2C, Abdennur N., Abraham S., Fudenberg G., Flyamer I.M., Galitsyna A.A., Goloborodko A., Imakaev M., Oksuz B.A., Venev S.V. Cooltools: enabling high-resolution Hi-C analysis in Python. Cooltools. - 2022.
8. Krueger F., Kreck B., Franke A., Andrews S.R. DNA methylome analysis using short bisulfite sequencing data//Nature Methods, 2012, Vol. 9, No. 2, P. 145-151.
9. Krueger F., Andrews S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications//Bioinformatics, 2011, Vol. 27, Bismark, No. 11, P. 1571-1572.
10. Ferragina P., Manzini G. Opportunistic data structures with applications // Proceedings 41st Annual Symposium on Foundations of Computer Science / 41st Annual Symposium on Foundations of Computer Science. – Redondo Beach, CA, USA: IEEE Comput. Soc, 2000. – C. 390-398.
11. Langmead B., Salzberg S.L. Fast gapped-read alignment with Bowtie 2//Nature Methods, 2012, Vol. 9, No. 4, P. 357-359.
12. Langmead B., Trapnell C., Pop M., Salzberg S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome//Genome Biology, 2009, Vol. 10, No. 3, P. R25.
13. Langmead B., Wilks C., Antonescu V., Charles R. Scaling read aligners to hundreds of threads on general-purpose processors//Bioinformatics, 2019, Vol. 35, No. 3, P. 421-432.
14. Pino F.M., Sukei E. fmorenopino/HeterogeneousHMM: First stable release of HeterogenousHMM. fmorenopino/HeterogeneousHMM. - Zenodo, 2020.
15. Shokoohi F., Stephens D.A., Bourque G., Pastinen T., Greenwood C.M.T., Labbe A. A Hidden Markov Model for Identifying Differentially Methylated Sites in Bisulfite Sequencing Data//Biometrics, 2019, Vol. 75, No. 1, P. 210-221.
16. Chen Y., Kwok C.K., Jiang H., Fan X. Detect differentially methylated regions using non-homogeneous hidden Markov model for bisulfite sequencing data//Methods, 2021, Vol. 189, P. 34-43.
17. Moreno-Pino F., Sükei E., Olmos P.M., Artés-Rodríguez A. PyHHMM: A Python Library for Heterogeneous Hidden Markov Models. PyHHMM / arXiv:2201.06968 [cs]. - arXiv, 2022.

4) Полученные результаты:

- В рамках сравнения двух подходов к построению библиотек для single-cell Hi-C (scHi-C) было показано, что PCR-протокол (основанный на биотин-меченых нуклеотидах, ферменте AluI и последующем обогащении химерных фрагментов) обеспечивает более

равномерное покрытие, меньшую долю артефактов и более высокую долю достоверных контактов, чем MDA-протокол, использующий амплификацию продуктами Phi29-полимеразы. Основные проблемы MDA-протокола были связаны с круговой переамплификацией и сменой матрицы, что приводило к множественным ложным контактам.

- Было разработано несколько метрик оценки качества scHi-C библиотек, включая:

— отношение (FF+RR) / (FF+RR+FR+RF) ориентаций прочтений;
— метрику «rings ratio» по данным прочтения длинных молекул (ONT);
— подсчёт числа фрагментов, участвующих более чем в 8 уникальных взаимодействиях как индикатор артефактов.

- В рамках метилирования было получено более 69% покрытия CpG позиций в геноме курицы при анализе отдельных ядер ооцитов. Средний уровень метилирования составил ~53.2%, что сопоставимо с соматическими тканями курицы (52.5–62.8%) и значительно выше, чем у сперматозоидов (~40.5%).
- Было показано, что метилирование хроматина в ооцитах птиц близко к соматическому, включая:

— гипометилированные CpG островки;
— гиперметилированные повторы и интергенные участки;
— отсутствие значимых различий между макро- и микрохромосомами.

- Сравнение стадий развития не выявило существенных различий в среднем уровне метилирования или в паттернах по геномным участкам, что говорит об отсутствии запрограммированного реметилирования при переходе между стадиями роста ооцита.
- Разработана HMM-модель сегментации метилирования, учитывающая расстояния между CpG сайтами и покрытие сайта. Она выделила три состояния: гипометилированное (<20%), промежуточное и гиперметилированное.
- С помощью модели было идентифицировано более 40 000 гипометилированных участков, из которых ~200 были специфичны для ооцитов. Эти регионы были обогащены промоторами и CpG-островками. Среди них отмечены ооцит-специфические профили метилирования у ключевых генов раннего развития:

— гипометилирование промотора NKX2-6;
— гипометилирование промотора TET2;
— гиперметилирование промотора KLF4.

5) Эффект от использования кластера в достижении целей работы:

Поскольку все данные получены высокопроизводительным секвенированием, то объёмы данных, которые необходимо проанализировать превосходят пределы типичных персональных компьютеров. Даже если найти достаточно оперативной памяти, то скорость вычислений превзойдёт все мыслимые пределы. Для решения таких задач использование вычислительного кластера с большим объёмом постоянной и оперативной памяти, а также с возможностью распараллеливания задач на сотни потоков – является определяющим успех решения моментом.