Сравнение пространственной структуры домена альфа-глобиновых генов в трех типах клеток G.gallus

Александра Галицына[1], Екатерина Храмеева[2,3], Сергей Ульянов[4]

[1] Московский Государственный Университет, Факультет Биоинженерии и Биоинформатики, Ленинские Горы, д.1, стр.73, Москва 119991, Россия
agalitzina@gmail.com
[2] Сколковский институт науки и технологий, ул.Новая, д.100, Сколково 143025, Россия
[3] Институт проблем передачи информации Российской академии наук, Большой Каретный переулок, д.19 стр. 1, Москва 127051, Россия
ekhrameeva@gmail.com
[4] Институт биологии гена Российской академии наук, ул. Вавилова, 34/5, Москва 119334, Россия
sergey.v.ulyanov@gmail.com

Аннотация Развитие методов определения конформации хромосом позволяет детально изучать взаимодействия участков хроматина в пространстве. Используя метод 5C, мы исследовали организацию области домена альфа-глобиновых генов курицы в трех типах клеток (в лимфоидных клетках, преэритробластах и индуцированных эритробластах), и выяснили, что хроматин организован в топологические домены (ТАДы) и разделяется на два компартмента активного и неактивного хроматина. В компартменте активного хроматина наблюдается более высокая экспрессия генов, а также большее количество меток ChIP-Seq архитектурного белка хроматина CTCF. Границы компартментов проходят по границам ТАДов и сохраняются между типами клеток и при индукции дифференцировки эритробластов. Домен альфа-глобинов расположен в компартменте активного хроматина. Его экспрессия отсутствует в лимфоидных клетках, идет на низком уровне в преэритробластах и на высоком уровне - в индуцированных эритробластах. Получены данные, свидетельствующие в пользу разрыхления хроматина при активации экспрессии генов альфа-глобинов.

Ключевые слова: фиксация конформации хромосом, 5C, альфа-глобиновые гены, ТАДы, компартменты хроматина

Изучение пространственной структуры хроматина с высоким разрешением стало возможным благодаря развитию высокопроизводительных методов фиксации конформации хромосом (С-методов). На данный момент существует множество вариантов С-методов, приспособленных для различных классов задач [1].

Активно исследуется вопрос влияния пространственной укладки хроматина на регуляцию экспрессии генов. Одна из популярных моделей исследования механизмов регуляции генов животных -- домен альфа-глобиновых генов [2]. В разных организмах домен устроен и функционирует схожим образом, и главной его

особенностью является специфичная экспрессия в эритроидных клетках (рис.1a).

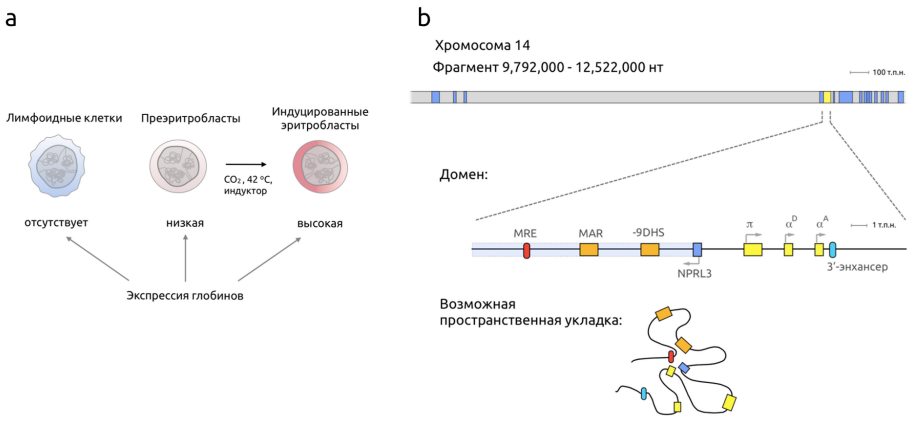

Рис. 1. (**a**) Модель исследования механизмов регуляции экспрессия альфа-глобиновых генов курицы. (**b**) Область альфа-глобинового домена курицы и его возможная пространственная укладка.

В частности, активно исследуется альфа-глобиновый домен курицы (*Gallus gallus*). Он расположен в 14 хромосоме на краю генной пустыни (области, не содержащей генов) и составляет в длину около 40 т.п.н. (рис.1b). Домен перекрывается с рамкой считывания гена NPRL3, направленной противоположно рамке считывания альфа-глобинов. Сам кластер глобиновых генов имеет в длину около 8 т.п.н. и включает 3 гена: эмбриональный альфа-глобиновый ген π, а также глобины α^D и α^A. Кроме того, домен содержит несколько регуляторных элементов, среди них: главный регуляторный элемент (MRE), расположенный в пятом интроне гена NPRL3, 3'-энхансер. Домен фланкирован участками прикрепления к ядерному матриксу (MAR) [3].

Исследования на лимфоидных и эритроидных клетках курицы с помощью метода 3C показали, что домен имеет сложную пространственную укладку, которая зависит от экспрессии генов альфа-глобинов и меняется с дифференцировкой эритроидных клеток [3].

Тем не менее, на данный момент неизвестно, как устроена и меняется структура домена альфа-глобиновых генов курицы в контексте окружающего хроматина. Исследование не только домена, но и окружающей области может расширить понимание строения хроматина и его связи с регуляцией экспрессии.

Метод фиксации конформации хромосом 5C (Chromosome Conformation Capture Carbon Copy) подходит для исследования области домена альфа-глобиновых генов [4], но ранее не применялся для клеток курицы. Он позволяет анализировать взаимодействия хроматина внутри некоторого региона генома. Его методика

(рис.2a) состоит в следующем: хроматин в ядрах исследуемых клеток сшивается с помощью формальдегида и нарезается рестриктазой, фрагменты лигируются и выделяются. С большей вероятностью лигируются фрагменты, сближенные в трехмерном пространстве ядра. К ним добавляются заранее сконструированные 5C-праймеры (рис.2b), комплементарные участкам вблизи сайтов рестрикции. Праймеры, которые отожглись на продукт лигирования, сшиваются стык в стык. На концах 5C-праймеров имеются универсальные последовательности, с помощью которых сшитые праймеры амплифицируются. Полученная ДНК отправляется на секвенирование. Чем большее количество раз представлен продукт лигирования двух фрагментов, тем ближе они расположены внутри ядра.

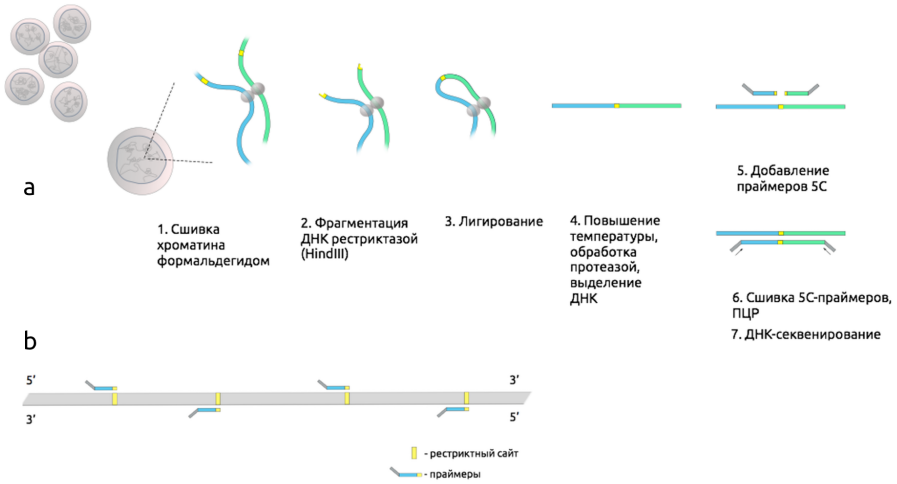

Рис. 2. (**a**) Схема эксперимента 5C. (**b**) Библиотека праймеров 5C.

В последние годы активно используется полногеномный C-метод HiC [5], для анализа которого разработаны разнообразные вычислительные приемы. Многие из них не применялись для данных 5C, тем не менее, это может помочь расширить знания о структуре хроматина исследуемой области.

В данной работе проанализированы данные 5C для области домена альфа-глобинов в трех типах клеток курицы: лимфоидных, преэритробластах и индуцированных эритробластах. Применены разнообразные вычислительные методы. Полученные данные свидетельствуют в пользу наличия топологических доменов (ТАДов) хроматина и разделения региона на компартменты активного и неактивного хроматина. Также получено подтверждение гипотезы о разрыхлении хроматина вокруг домена альфа-глобинов при активации их экспрессии.

1. Методы

Для исследования были взяты результаты уже готовых экспериментов: Illumina секвенирования 5C-данных, RNA-Seq и ChIP-Seq сайтов связывания архитектурного белка хроматина CTCF, для трех типов клеток *G.gallus*: лимфоидных клеток DT40, клеток преэритробластов HD3 и индуцированных эритроблатов HD3ind.

Индукция преэритробластов проводилась с помощью добавления вещества-индуктора и выдерживания клеток при $42°$C [6]. Планирование библиотеки 5C-праймеров проводилось с помощью сервиса My5C [7]. Праймерами был покрыт фрагмента генома *G.gallus* 9,792,000-12,552,000 нт на 14 хромосоме. Эксперимент 5C проводился по стандартной методике [4].

Для картирования данных RNA-Seq использовалась программа tophat2. Для картирования данных 5C была использована программа bowtie2 [14]. Также было протестировано картирование в помощью простого поиска в библиотеке праймеров. Коэффициент корреляции Пирсона между результатами разных методов картирования был больше 0.97 для каждого эксперимента.

Для каждого эксперимента 5C имелись данные по двум повторностям. Коэффициент корреляции Пирсона между повторностями после этапа картирования был более 0.9 для каждого типа клеток. Из рассмотрения были удалены праймеры, взаимодействия для которых были близки к нулю хотя бы для одного типа клеток. Полученные данные бинировались по окну в 30 т.п.н. Результат был представлен в виде матрицы взаимодействий бинов. Диагональные элементы удалялись из рассмотрения.

Далее производилась итеративная коррекция [8] с приведением количества взаимодействий в разных экспериментах к общему среднему. После этого этапа проводилось объединение повторностей простым сложением взаимодействий для соответствующих пар бинов. Для уменьшения шума и лучшей визуализации проводилось сглаживание -- усреднений взаимодействий соседних бинов.

Для сравнения взаимодействий между типами клеток использовался подсчет относительной разности (коэффициентов Жаккара, Jaccard) по формуле:$J(a,b) = \frac{a-b}{a+b}$, где a и b -- значения взаимодействий в сравниваемых экспериментах для заданной пары бинов.

Данные после этапа сглаживания визуализированы с помощью теплокарт взаимодействий исследуемого региона. Для карты каждого эксперимента количество взаимодействий увеличивалось к диагонали. Вдоль диагонали наблюдались топологические домены (ТАДы) [9] -- области с более высоким количеством взаимодействий, чем окружение. Автоматизация поиска ТАДов осуществлялась с помощью алгоритма Armatus [10], реализованном в библиотеке Greendale для Python [11]. Параметр работы алгоритма выбирался так, чтобы полученные ТАДы наи-

лучшим образом соответствовали их визуальному восприятию.

Для обнаружения декомпактизации в области альфа-глобинового домена проводился тест на изменение частоты взаимодействий внутри ТАДов [12]. Для этого подсчитывалось количество бинов ТАДа, которые уменьшили количество взаимодействий между двумя типами клеток. Значимость уменьшения в ТАДе оценивалась с помощью следующего теста. Брались бины внутри всех ТАДов, из них 10000 раз случайным образом формировались ТАДы того же размера, что и анализируемый. Каждый раз считалось количество бинов, уменьшивших взаимодействия. По тому, в какую квантиль распределения попадает наблюдаемая величина, можно судить о значимости уменьшения взаимодействия.

Поиск компартментов хроматина выполнялся с помощью PCA [13]. Для каждого эксперимента строилась зависимость количества взаимодействий от расстояния. По ней строилась матрица ожидаемых взаимодействий, на которую нормировалась наблюдаемая матрица. PCA-анализ проводился по матрице коэффициентов корреляции Пирсона для нормированной на ожидание матрицы.

Поиск компартментов проводился для сглаженных и несглаженных данных. Наибольшую долю дисперсии объясняли две или три первых главных компоненты. Для сглаженных данных первая компонента делилась на две области, соответствующие значениям с противоположными знаками. Для несглаженных данных такое деление имела вторая компонента. Такие области с противоположными знаками интерпретируются как два компартмента хроматина [13]. Границы компартментов одинаковы между клеточными линиями, а также между сглаженными и несглаженными данными.

2. Результаты

В работе были использованы результаты следующих экспериментов для трех типов клеток курицы (лимфоидных клеток, преэритробластов и индуцированных эритробластов): 5C, ChIP-Seq за структурный белок хроматина CTCF и RNA-Seq.

Было выполнено простроение карт взаимодействий 5C (картирование, фильтрация, бинирование, коррекция, объединение реплик, сглаживание). По ним произведен анализ главных компонент и поиск ТАДов. Также произведено картирование данных RNA-Seq. Данные ChIP-Seq были предоставлены в виде готовых пиков.

Результат наложения полученных разметок для лимфоидной линии представлен на рис.3, для преэритробластов -- на рис.4, для индуцированных эритробластов -- рис.5.

Примечательно, что во всех случаях границы комартментов проходят по границам ТАДов. В одном из компартментов сосредоточено большинство меток CTCF и участки с активной экспрессией. Согласно терминологии, введенной в [13], это компартмент A активных и потенциально активных генов. Пики CTCF, возможно, означают образование петель хроматина в области активных генов. В другом

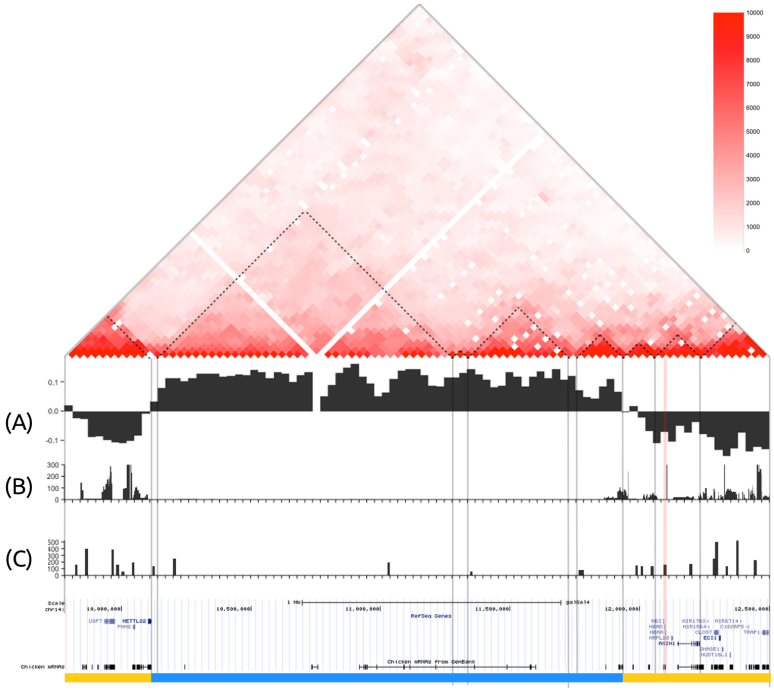

Рис. 3. Наложение полученных разметок для лимфоидной линии (DT40). Приведена сглаженная карта, ТАДы (Armatus $\gamma = 0.09$ по сглаженной карте), первая компонента PCA (A), данные RNA-Seq (B) и ChIP-Seq (C). Ниже представлена разметка генов и мРНК из UCSC. Розовой вертикальной прямой обозначено положение домена альфа-глобинов. Внизу желтые прямоугольники означают компартмент А, синий прямоугольник -- компартмент В. Пустые строка и столбец карты - бины с отсутствующими данными.

компартменте экспрессия практически не происходит, мало меток CTCF. Это компартмент В или генная пустыня. Интересно, что в компартменте В находятся более крупные и рыхлые ТАДы, по сравнению с компартментом А.

Домен альфа-глобинов попадает в компартмент А, в центр ТАДа для эритроидных клеток и практически на границу ТАДов в лимфоидных клетках. Экпрессия с домена глобиновых генов отсутствует для лимфоидных клеток (рис.3В), идет на низком уровне для преэритробластов (рис.4В) и идет активно для индуцированных эритробластов (рис.5В).

Результат PCA-анализа для сглаженных и несглаженных данных свидетельствует о том, что компартменты и их границы сохраняются между типами клеток (рис.6).

Сравнения карт взаимодействий между типами клеток с помощью коэффициента Жаккара представлены на рис.7-8. В целом, при сравнении между типами

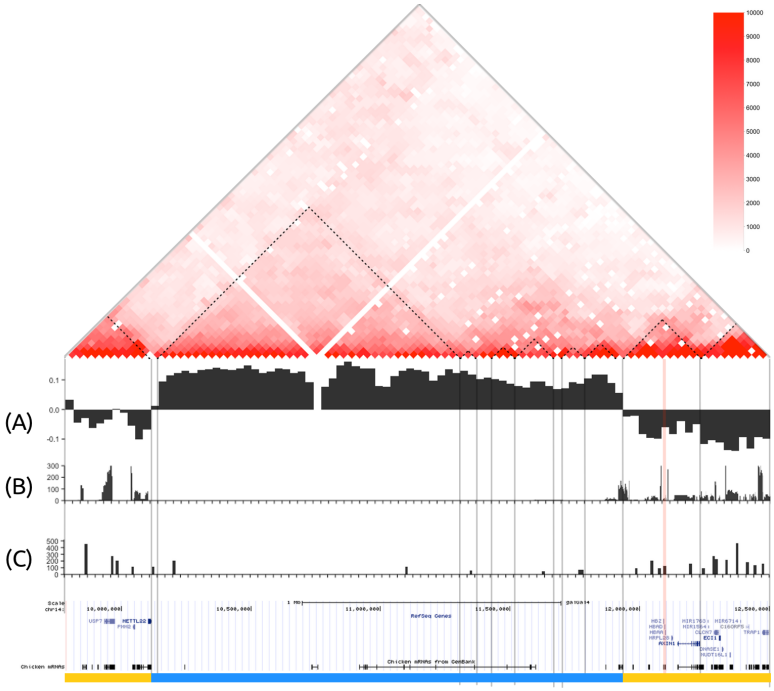

Рис. 4. Наложение полученных разметок для преэритробластов (HD3). Получение, обозначения и подписи те же, что и на рис.3.

клеток практически не наблюдается четких изменений количества контактов по ТАДам.

Следует отметить различие в области 5'-карты при сравнении лимфоидных клеток и преэритробластов (рис.7), сопряженное в эритроидной линии с понижением экспрессии и значением первой компоненты PCA, более смещенной к компартменту B.

Можно заметить, что в индуцированных клетках по сравнению с неиндуцированными (рис.8) наблюдается экспрессия длинного транскрипта в 3'-области генной пустыни. В аннотации UCSC на этом месте нет генов (ср. рис.5). Возможно, это экспрессия неаннотированного гена. Интересно, что значение первой компоненты на 3'-конце при этом становится более смещенным в сторону A-компартмента.

Для сравнения ТАДов между типами клеток использовались разметки, полученные алгоритмом Armarus при параметре 0.09. Результат приведен на рис.9. Плотность взаимодействий в ТАДе, содержащем глобиновый домен, понижается в индуцированных эритробластах по сравнению с преэритробластами.

Это подтверждается тестом на понижение количества взаимодействий. Всего в ТАДе с глобинами содержится 104 бина, из которых количество бинов, уменьшивших взаимодействия, равняется 64. Ожидаемое количество бинов, уменьшивших

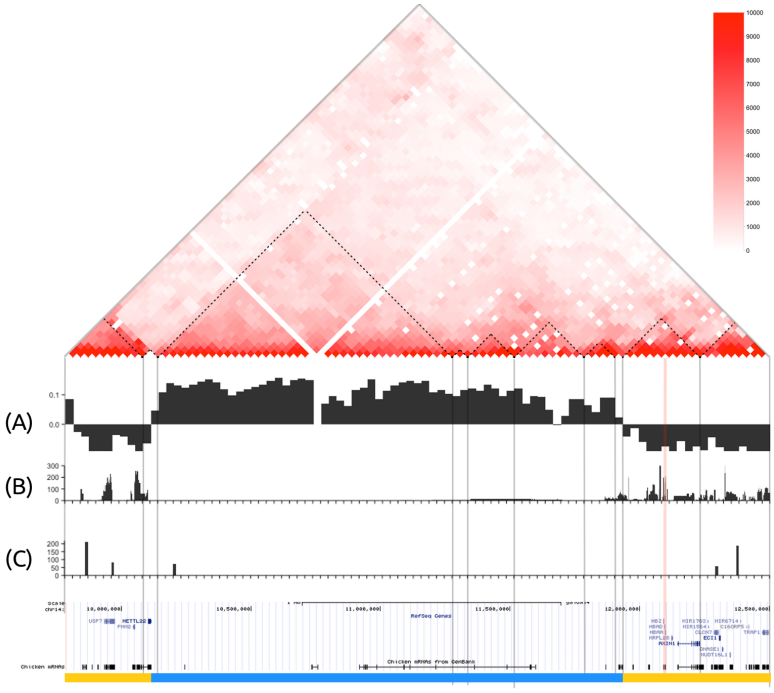

Рис. 5. Наложение полученных разметок для индуцированных эритробластов (HD3ind). Обозначения и подписи те же, что и на рис.3.

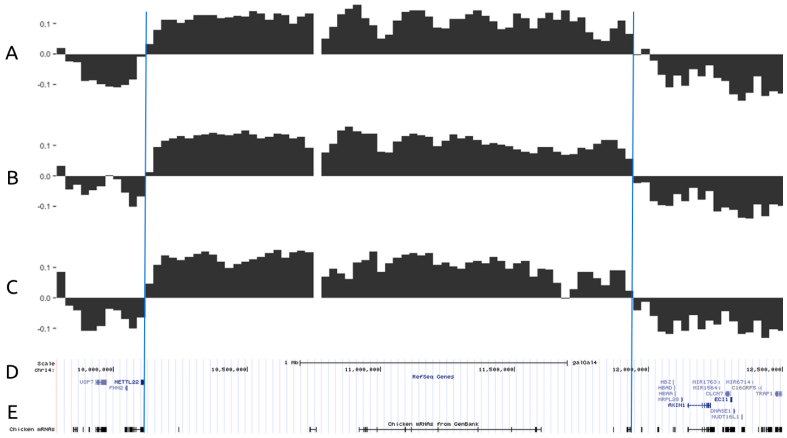

Рис. 6. Компартменты для (A) лимфоидных клеток, (B) преэритробластов, (C) индуцированных эритробластов. Разметка генов (D) и мРНК (E) курицы из UCSC.

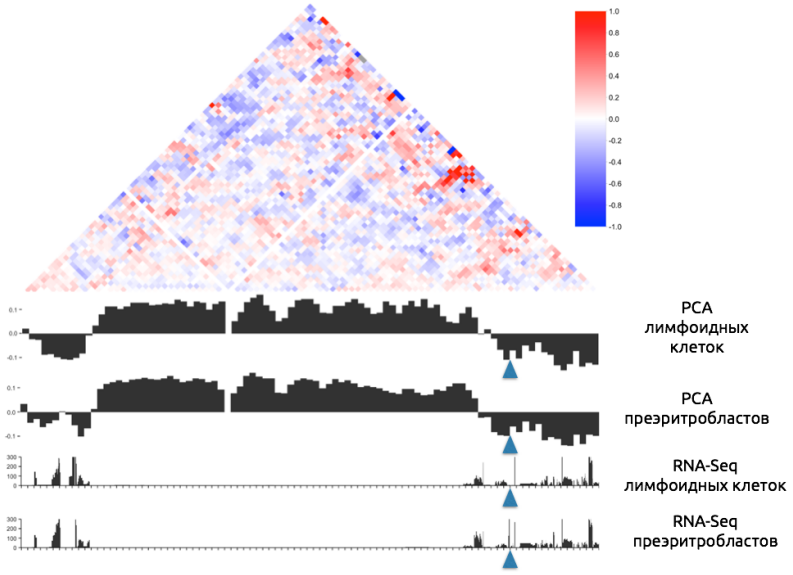

Рис. 7. Карта Жаккара для сравнения лимфоидных клеток и преэритробластов. Синей стрелкой показано положение домена альфа-глобинов.

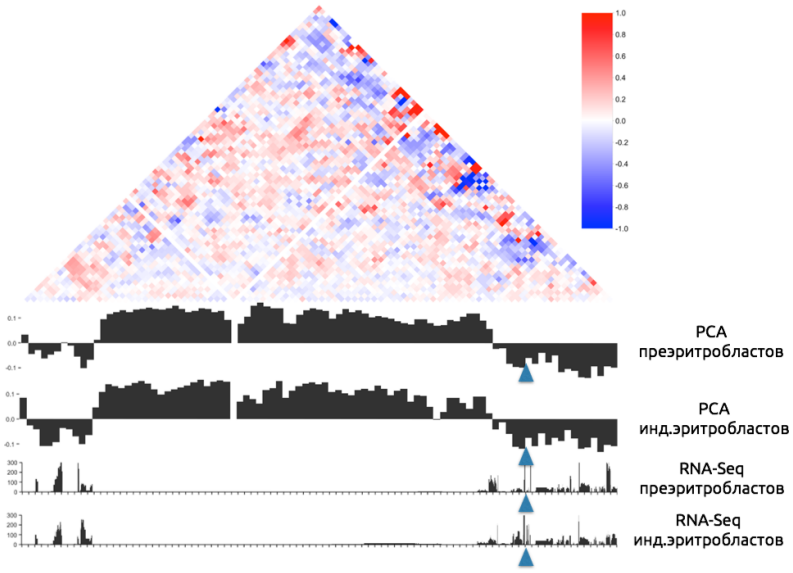

Рис. 8. Карта Жаккара для сравнения преэритробластов и индуцированных эритробластов.

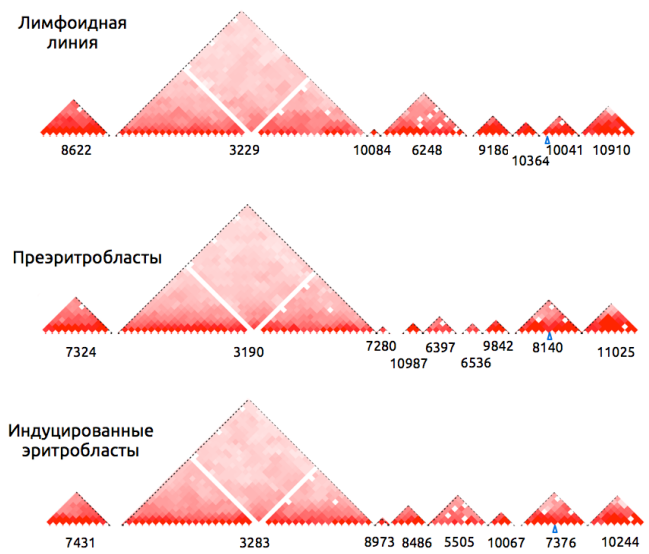

Рис. 9. Сравнение плотности взаимодействий по ТАДам сглаженных карт (Armatus, $\gamma = 0.09$)

взаимодействия, при случайном формировании ТАДа -- около 52. Наблюдаемая величина попадает в 0.01 верхнюю квантиль распределения. Это свидетельствует в пользу гипотезы о декомпактизации домена альфа-глобинов при активации их экспрессии.

Сравнение плотностей между лимфоидными и эритроидными клетками затруднено, так как ТАДы с глобинами в них имеют разные размеры. Тест на понижение количества взаимодействий может быть применен только для ТАДов равных размеров.

3. Обсуждение

В работе проведен анализ данных 5C. В частности, использовались методы анализа, ранее применявшиеся только для полного генома или отдельных хромосом.

Тем не менее, эти методы оказались переносимыми и на небольшой участок хромосомы.

Этапы картирования, фильтрации, бинирования и сглаживания данных являются традиционными для 5C.

Выбор программы картирования bowtie2 может быть спорным, так как наилучшим образом она работает с длинными прочтениями более 100 нт в длину [14]. Тем не менее, применение метода простого поиска праймеров в данных практически воспроизвело результаты bowtie2. Сходимость двух принципиально разных

методов свидетельствует о том, что любой из них подходит для картирования. Тем не менее, выбран метод bowtie2 как более традиционный.

Для уменьшения систематической погрешности, происходящей от неравномерной представленности взаимодействий праймеров, был использован этап итеративной коррекции. Кроме неё, был протестирован метод BAC-нормировки [4], но он внес дополнительные погрешности и не исправил существующие (данные не приведены).

Этап сглаживания необходим для уменьшения шума в данных 5C и улучшения восприятия. Однако, большинство результатов воспроизводится и на несглаженных картах.

Поиск границ ТАДов осуществлялся с помощью алгоритма Armatus, с выбором параметра 0.09. В работе были протестированы другие параметры и методы. Алгоритм Armatus давал результаты, наилучшим образом согласующиеся с визуальным восприятием ТАДов, а также воспроизводимые при многих значениях параметров.

Примечательно, что размер найденных ТАДов согласуется с данными о ТАДах для других организмов -- от нескольких сотен Kb до Mb.

Неожиданным и противоречащим данным литературы [12] является тот факт, что ТАДы меняют свои границы между типами клеток. В наших данных изменения происходят в правой половине карты, где имеется наибольшее количество бинов с отсутствующими данными (рис.9). Причина нестабильности этой области может быть в нехватке данных, но нельзя исключать возможность биологических причин наблюдаемого эффекта.

Поиск компартментов с помощью PCA позволил разделить область на компартменты активного и неактивного хроматина, согласующимся по размерам и характеристикам с данными из литературы [15].

На настоящий момент такие компартменты были открыты в клетках человека и мыши [15]. Полученные данные подтверждают их существование и у курицы.

Домен альфа-глобинов попадает в компартмент активного и потенциально активного хроматина во всех трех типах клеток. Его экспрессия меняется между клеточными линиями в соответствии с ожиданием.

Домен попадает в центр крупного ТАДа в эритроидных клетках. Этот ТАД распадается на два более мелких в лимфоидной линии. Это наблюдение не согласуется с данными из литературы, что на границах ТАДов наблюдается повышенная активность экспрессии. Причин этому может быть несколько. Во-первых, это может быть следствием эксперимента и обработки 5C (например, можно подобрать такой параметр алгоритма поиска, что ТАД с глобинами не будет распадаться). Во-вторых, это может быть биологическое явление, объясняемое специфической петлевой структурой.

Примечательно, что плотность взаимодействий в ТАДе с глобинами уменьшается при индукции эритробластов. Это подтверждается тестом на изменение частоты взаимодействий внутри ТАДов и уменьшением плотности взаимодействий по абсолютной величине (данные не приведены). Сравнение с лимфоидными клетками затруднено, так как в них ТАД с глобинами меняет свои границы и размеры.

Список литературы

1. E. Wit, W. Laat, *A decade of 3C technologies: insights into nuclear organization*, GENES and DEVELOPMENT 2012, 26:11–24
2. D. Bau et al., *The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules*, Nat Struct Mol Biol. 2011, 18(1): 107–114
3. A.A. Galrilov, S.V. Razin, *Spatial configuration of the chicken alpha-globin gene domain: immature and active chromatin hubs*, Nucleic Acids Research 2008, 36(14): 4629–4640
4. J. Dostie et al., *Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements*, Genome Research 2006, 16: 1299–1309
5. N.L. Berkum et al., *Hi-C: A Method to Study the Three-dimensional Architecture of Genomes*, Journal of Visualized Experiments 2010, 39: e1869
6. A.A. Gavrilov, S.V. Razin, O.V. Iarovaia, *C-methods to study 3D organization of the eukaryotic genome*, Biopolymers and Cell 2012, 28(4):245–251
7. B.R. Lajoie et al., *My5C: webtools for chromosome conformation capture studies*, Nature Methods 2009, 6(10): 690–691
8. M. Imakaev et al., *iterative correction of hi-c data reveals hallmarks of chromosome organization*, Nature methods 2012, 9(10): 999-1003
9. J.R. Dixon et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions*, Nature 2012, 485: 376-380
10. D. Fillipova et al., *Identification of alternative topological domains in chromatin*, Algorithms for Molecular Biology 2014, 9:14
11. https://bitbucket.org/nvictus/greendale/
12. S.S. Rao et al., *A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping*, Cell 2014, 159: 1–16
13. E. Lieberman-Aiden et al., *Comprehensive mapping of long range interactions reveals folding principles of the human genome*, Science. 2009, 326(5950): 289–293
14. http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
15. J.R. Dixon et al., *Chromatin architecture reorganization during stem cell differentiation*, Nature 2015, 518: 331-336