МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Факультет естественных наук
Кафедра цитологии и генетики

Направление подготовки «Биология» (06.03.01)

КУРСОВАЯ РАБОТА БАКАЛАВРА

Мякинькова Ивана Олеговича

Тема работы: Характеристика нуклеотидных последовательностей в основаниях дальних хроматиновых петель у комаров рода *Anopheles*

«К защите допущена»

Заведующий кафедрой,

……/……..

(фамилия , И., О.) / (подпись)

«……»……20…г.

Научный руководитель

Фишман Вениамин Семёнович

к. б. н., в. н. с. ИЦиГ СО РАН

……/……..

(фамилия , И., О.) / (подпись)

«……»……20…г.

Новосибирск, 2022

Список сокращений

т. п. н. — тысяча пар нуклеотидов

м. п. н. — миллион пар нуклеотидов

ТАД — топологически ассоциированный домен

Hi-C — high-throughput chromosome conformation capture

H3K27me3 — триметилирование двадцать седьмого лизина гистона H3

H3K27ac — ацетилирование двадцать седьмого лизина гистона H3

H3K9me3 — триметилирование девятого лизина гистона H3

ChIP-seq — иммунопреципитация хроматина и высокоэффективное секвенирование ДНК

Оглавление

1. Введение

Геном эукариотических клеток — это сложная многоуровневая система со строгой иерархичностью. В ходе всего клеточного цикла он задействуется во

множестве молекулярных процессов, в том числе процессах матричного синтеза. Для корректного и своевременного воспроизведения генетической информации важно, как хроматин расположен внутри ядра. Совокупность закономерностей такого расположения называется пространственной организацией генома. На сегодняшний день трёхмерная геномика сосредоточена на млекопитающих (на человеке и мышах), а строению хроматина насекомых уделяется меньше внимания. Современные представления о 3D-геноме насекомых получены в основном из работ с представителями рода *Drosophila*.

Малярийные комары бросают вызов современному здравоохранению как активные и вездесущие векторы патогенных вирусов, бактерий и простейших, самыми опасными из которых являются малярийные плазмодии. Возрастает адаптивная способность комаров и, как следствие, широта их миграции.

Изучение трёхмерной организации хроматина малярийных комаров актуально, поскольку понимание 3D-организации их генома позволит управлять молекулярными механизмами и физиологией комаров и контролировать распространение ими заболеваний. Развитие 3D-геномики комаров также может пролить свет на противоречивую природу доменов хроматина у плодовых мушек — с одной стороны, их изменения далеко не всегда ведут к изменениям в генной экспрессии [1], в то же время они очень консервативны [2].

В работе Lukyanchikova *at al*. (2022) [3] в геномах пяти видов комаров рода *Anopheles* были обнаружены необычайно длинные хроматиновые петли (до 31 м. п. н.). Их образование не может быть объяснено ни одним известным на сегодняшний день механизмом. Детальный анализ последовательностей, которые лежат в основаниях этих петель, не проводился. Поэтому представляет интерес структура этих последовательностей и их сходство у разных видов, так как это может привести ближе к пониманию механизмов формирования необычных петель и устройства генома анофелесов в целом.

В связи с этим в курсовой работе была поставлена цель выделить и охарактеризовать функциональные элементы генома, лежащие в основаниях длинных хроматиновых петель у комаров рода *Anopheles*.

Задачи, которые необходимо выполнить для достижения цели:

1. Используя данные секвенирования третьего поколения, улучшить существующие сборки геномов комаров рода *Anopheles*.

2. В усовершенствованных геномах выявить и охарактеризовать повторяющиеся последовательности в основаниях дальних петель.
3. Определить консервативные регионы в основаниях дальних петель у представителей разных видов анофелесов.

2. Литературный обзор

2.1. Общие принципы трёхмерной организации генома

Трёхмерная (пространственная, 3D-) организация генома — набор закономерностей, согласно которым молекулы ДНК располагаются в пространстве клеточного ядра. В последние годы стало известно, что трёхмерная организация связана не только с компактизацией длинных молекул ДНК в ограниченном внутриядерном пространстве, но и с контролем важных молекулярно-генетических процессов: репликации [4], транскрипции [5] и репарации [6] ДНК.

3D-организация интерфазного генома иерархична и включает несколько уровней, отличающихся по масштабу и структуре [7]:

1. Хромосомная территория — часть объёма клеточного ядра, занятая одной хромосомой (молекулой ДНК). Это самый масштабный уровень 3D-организации. Наличие хромосомных территорий легко продемонстрировать при помощи экспериментов по флуоресцентной гибридизации *in situ*.
2. Компартмент — область ядра, занятая хроматином с одним типом эпигенетических меток — сходным набором гистоновых модификаций, паттерна метилирования и уровня генной экспрессии. По размерам меньше хромосомных территорий, но могут содержать участки разных хромосом. Наиболее простая классификация включает два типа компартментов: A-компартмент (эухроматин) содержит активно экспрессирующиеся гены, B-компартмент (гетерохроматин) — «молчащие» гены и нетранскрибируемые участки хроматина без активных геномных элементов. Гетерохроматин находится на периферии ядра (около ламины) и в ядрышках, эухроматин заполняет остальное пространство. Компартменты одного типа кластеризуются. Размер компартмента может составлять до 1 м. п. н. Были открыты раньше хромосомных территорий с помощью дифференциальной окраски хромосом и световой микроскопии. С появлением методов анализа контактов хроматина с высоким разрешением представления о компартментах ядра существенно расширились, о чем будет более подробно сказано ниже.
3. Топологически ассоциированный домен (ТАД) — непрерывный участок генома, инсулированный от окружения и содержащий локусы, часто контактирующие друг с другом. ТАДы обнаружены у большого количества видов, хотя механизмы их образования отличаются. Из-за формирования ТАДов последовательности, расположенные на большом линейном расстоянии в геноме, могут взаимодействовать, благодаря чему удалённые от промоторов энхансеры способны управлять транскрипцией генов. ТАДы, образованные механизмом экструзии петли (см. далее), не статичны, но представляют собой динамичные структуры. Размер ТАДа составляет от сотен т. п. н. до 1 м. п. н. Были открыты с помощью методов захвата конформации хромосом (C-методов), в том числе Hi-C, описанного ниже.

4. Хроматиновые петли (петлевые домены) образуются в результате пространственного сближения удалённых друг от друга последовательностей ДНК — оснований петель (англ. *anchors*). Участок ДНК, фланкированный основаниями, и образует петлю. Петли могут существовать конститутивно, либо в определённом клеточном типе (например, образованные в результате сближения промотора гена β-глобина со своим энхансером [8]), могут образовываться разово в течение небольшого промежутка времени (петли, образующиеся во время V(D)J-рекомбинации в лимфоцитах [9]). Различные биологические механизмы, которые могут лежать в основе формирования петель, будут описаны ниже.
5. Нуклеосомный уровень описывает расположение гистонов на ДНК, частоту и плотность их посадки. Нуклеосомная организация связана с гистоновым кодом — закономерностью распределения химических модификаций (меток) гистоновых белков, например, триметилирования или ацетилирования двадцать седьмого лизина гистона H3 (H3K27me3 или H3K27ac). Гистоновые метки — ацетилирование, метилирование, гликозилирование, убиквитинилирование и многие другие — определяют плотность расположения нуклеосом и в принципе их наличие на ДНК, что влияет на доступ к ней транскрипционных факторов. Организация хроматина на уровнях компартментов и нуклеосом связана: для более компактной упаковки ДНК (гетерохроматина) характерны одни гистоновые метки (например, триметилирование девятого лизина гистона H3, H3K9me3), для менее компактной (эухроматина) — иные (например, ацетилирование гистонов всегда ассоциировано с A-компартментом).

Одним из наиболее популярных методов исследования пространственной организации хроматина в наше время стал метод high-throughput chromosome conformation capture (Hi-C) [10]. Он позволяет количественно оценить частоту физических контактов всех участков генома друг с другом в пространстве. Метод основан на обработке генома формальдегидом, образующим ковалентные связи белок-белок и ДНК-белок. В результате участки генома, расположенные близко друг ко другу, с большей вероятностью окажутся ковалентно связаны. Затем геном фрагментируют, ДНК лигируют, так что полученные молекулы состоят из участков генома, которые находились в близости друг ко другу. Из этих фрагментов собирают библиотеки для секвенирования нового поколения. Полученные в результате секвенирования прочтения будут выравниваться на последовательность референсного

генома химерно, то есть отдельные части прочтений могут быть выровнены на отстоящие друг от друга последовательности ДНК. Подсчитав количество химерных прочтений, можно определить частоту, с которой данные участки генома оказывались сближены в пространстве ядра. Частоты вносятся в симметричную матрицу (Hi-C-карту), по осям которой установлены геномные координаты; если присвоить численным значениям частот интенсивность цвета, получится тепловая карта пространственных контактов, удобная для визуальной интерпретации.

2.2. Механизмы формирования хроматиновых петель

На сегодняшний день известно несколько механизмов формирования петель хроматина у эукариот.

2.2.1. Экструзия петли с барьерными элементами

Хроматиновые петли млекопитающих были обнаружены с помощью C-методов (5C, Hi-C) [11]. Многие из петлевых взаимодействий консервативны у млекопитающих [12]. У большинства петель хроматина млекопитающих на границах располагается CCCTC-связывающий фактор (CTCF) вместе с одним из комплексов структурной поддержки хромосом (SMC) — когезином [13]. CTCF — повсеместно экспрессируемый белок с мотивом из 11-и цинковых пальцев, связывающийся с последовательностью CCGCGNGGNGGCAG (если она не метилирована) и способный образовывать ди- и мультимеры [14]. Границы петель вовлечены в сильное пространственное взаимодействие, которое на карте контактов Hi-C отображается как яркие красные точки (угловые пики) на вершинах треугольников. Сам же треугольник контактов свидетельствует о сближении участка хроматина между основаниями с CTCF. Сайты посадки CTCF в основании петель расположены в конвергентной ориентации (направлены друг навстречу другу, →←). Если сайты посадки расположены в иной ориентации (дивергентной ←→, или сонаправленно, →→), петля не формируется. Эти наблюдения объясняют механизмом экструзии петли (англ. *loop extrusion*) [15], согласно которому белковый комплекс когезин связывается с ДНК и начинает «протягивать» её до тех пор, пока не диссоциирует или не встретит на пути конвергентно ориентированные CTCF. В последнем случае комплекс останавливается, и сайты посадки CTCF становятся основаниями петли. Когезин представляет собой белковый комплекс, основной элемент которого — «кольцо» из белков SMC1, SMC3 и

RAD21. Оно содержит АТФазные домены, за счёт которых может перемещаться по ДНК, причём как топологически (ДНК находится внутри кольца, как продетая через него), так и нетопологически, то есть просто скользя по ДНК, не заключая её внутрь. В соответствии с этим, при анализе иммунопреципитации хроматина (ChIP-seq) сигналы от когезина располагаются более внутренне по отношению к петле, чем сигналы от CTCF [15]. Эта модель подтверждается последними молекулярно-генетическими исследованиями. ТАДы, таким образом, образуются из-за невозможности когезином дальше протягивать петлю — он останавливается на некоторых конвергентных сайтах CTCF (не все сайты CTCF, даже конвергентно ориентированные, способны становиться границами ТАДов). Размеры ТАДов при этом не превышают 1—2 м. п. н.

Петли не формируются, если когезин не связан с хроматином или не остановился на конвергентно расположенных CTCF. Деплеция когезина, CTCF либо фактора загрузки когезина NIPBL ведёт к разрушению ТАДов [16]. С другой стороны, удаление фактора разгрузки когезина WAPL (он раскрывает когезиновое кольцо) увеличивает частоту контактов CTCF-оснований и содержимого их петель [17], самих петель при этом становится меньше, они удлиняются, геном в целом компактизуется. Если обратить деплецию когезина в клетке, петли моментально восстановятся . Следовательно, связь между когезином и CTCF находится в равновесии: с одной стороны, когезин диссоциирует или удаляется фактором WAPL, с другой — устанавливается фактором NIPBL, причём чем дольше когезин связан с CTCF, тем ярче становится угловой пик на Hi-C-карте. Описанные факты подтверждают центральную роль когезина в формировании хроматиновых петель и косвенно подтверждают механизм экструзии петли. Формирование петель является динамическим процессом: один и тот же ТАД, обнаруженный через Hi-C на клеточной популяции, выглядит по-разному в каждой из этих клеток при single-cell Hi-C, так как в отдельных клетках когезин находится на разных стадиях выпетливания ДНК или вовсе от неё отсоединён [18].

Механизм экструзии петли был показан напрямую *in vitro* в экспериментах на дрожжах [19]. С помощью высокоразрешающей микроскопии продемонстрировали комплекс когезина, перемещающийся по молекуле ДНК и формирующий из неё петлю.

Использование флуоресцентной гибридизации *in situ* продемонстрировало наличие неких петлевых доменов как в клетках дикого типа, так и в клетках с деплецией когезина, но в первом случае их границы располагались строго на сайтах посадки CTCF, а во втором — случайным образом [20]. Следовательно, существуют

иные механизмы, формирующие петлевые домены в отсутствие когезина, либо петли формируются спонтанно.

2.2.2. Yin Yang 1

При дифференцировке мышиных эмбриональных стволовых клеток (mESC) в нервные прогениторные клетки (NPC) было замечено значительное уменьшение количества CTCF, связанного с ДНК. Также при этом уменьшалась частота пространственных взаимодействий промоторов и энхансеров, характерных для mESC — так отражается снижение потенциала к дифференцировке в разные клеточные типы. В то же время наблюдалось увеличение связывания с ДНК белка Yin Yang 1 (YY1). Он часто связывал контактирующие в пространстве гены и энхансеры, специфичные для NPC, при этом отсутствовал на тех, что не были сближены. Нокдаун гена *Yy1* приводил к исчезновению многих пространственных взаимодействий в геноме [21].

Прежде о YY1 было известно, что это транскрипционный фактор, которым обогащены петлевые взаимодействия в культурах клеток человека и который необходим для нормальной нейральной дифференцировки. Этот белок экспрессируется повсеместно, связывается с мотивом 5’-CCGCCATNTT-3’, который распознаёт цинковыми пальцами (но только если мотив гипометилирован), способен образовывать гомодимеры, и без него нарушается эмбриональное развитие. Этими характеристиками он очень схож с CTCF, он является как бы его «двойником».

Таким образом, YY1 является очень хорошим кандидатом на роль архитектурного белка хроматина. Не ясно, каким именно механизмом он опосредует формирование хроматиновых петель, предлагается три варианта: (1) два белка связывают ДНК в разных местах, а затем соединяются в гомодимер; (2) YY1 связывает ДНК, а затем формирует гетеродимер с CTCF, также связанным ДНК; (3) YY1 является барьерным элементом в механизме экструзии петли, который останавливает перемещение по ДНК когезина или другого экструдера.

Работа Weintraub *at al.* (2017) [22], проведённая на mESC, демонстрирует, что YY1, во-первых, связывает последовательности активных энхансеров и промоторов, во-вторых, образует гомодимеры. Ряд экспериментов, проведённых в этой работе, подтверждает роль YY1 в организации пространственного взаимодействия ДНК и образовании петель. Так, при делеции мотива связывания YY1 в промоторе гена *Raf1* наблюдалось уменьшение связанного с промотором белка YY1, снижение частоты

контактов этого промотора со своим энхансером и уменьшение количества транскриптов этого гена. Если установить dCas9, соединённый с YY1, рядом с промотором, лишённым мотива связывания YY1, частота контактов этого промотора с энхансером вырастет, а транскрипция гена под этим промотором — увеличится (такой эксперимент можно считать «спасением фенотипа»). Деплеция YY1 вызывает значительное (до 50 %) снижение частоты контактов между промоторами и энхансерами, на которых он был обнаружен, а также снижение экспрессии генов (на 30—40 %), с промоторами которых был связан. При этом через некоторое время после деплеции количество белка восстанавливается, и частоты промотор-энхансерных взаимодействий повышаются. Наконец, эта работа показывает сближение белком YY1 молекул ДНК *in vitro*. Линейные последовательности ДНК, содержащие мотивы связывания YY1, лигировались и закольцовывались чаще, чем без неё, поскольку белок сближал концы молекул в пространстве, увеличивая шанс их встречи с лигазой. Авторы предполагают, что YY1-опосредованное энхансер-промоторное петлеобразование — особенность, присущая всем млекопитающим.

Важное наблюдение в работе [21] в том, что в NPC взаимодействия YY1-YY1 наблюдались внутри петель, образованных белком CTCF ранее, на стадии mESC. В целом сайты посадки YY1 в 30 % случаев располагались рядом с сайтами посадки CTCF — возможно, два эти белка работают в паре, совместно регулируя пространственную организацию хроматина у млекопитающих.

Существует много других транскрипционных факторов с цинковыми пальцами, способных к гомо- и гетеродимеризации. Возможно, какие-то из них в будущем также будут распознаны как архитектурные белки генома.

2.2.3. MyoD

Белок животных MyoD также вызывает формирование хроматиновых петель в геноме. Прежде он был известен как стадиеспецифичный транскрипционный фактор, экспрессирующийся только в сомитах и необходимый для дифференцировки миоцитов [23]. Недавнее исследование [24] показало, что регуляцию клеточного развития этот фактор выполняет посредством реорганизации 3D-генома. Прежде было известно, что белки MyoD и Myf5 совместно регулируют экспрессию генов мышечной дифференцировки, связываясь с мотивом E-box по всему геному. Мыши без одного из этих факторов развивают нормальную мускулатуру, но миогенез полностью прекращается при нокаутировании их обоих [25]. На миоцитах с нокаутом по MyoD

было показано, что этот фактор вместе с CTCF участвует в инсуляции границ ТАДов. Это привело к выводу, что MyoD, как и CTCF, может контролировать формирование хроматиновых петель. Действительно, значительная часть оснований хроматиновых петель в миоцитах обогащена мотивом E-box и связана одновременно с CTCF и с MyoD. Нокаут последнего приводил к существенному ослаблению петель и с MyoD, и с CTCF в основаниях — значит, именно MyoD определяет, где будут располагаться петли в миоцитах; его возвращение в клетку восстанавливало петли. Удаление последовательности E-box из оснований, образованных под «руководством» MyoD, приводило к исчезновению петель. Таким образом, MyoD является белком, сближающим свои консенсусные последовательности и образующим хроматиновые петли. Примечательно, что они образуются преимущественно в миоцитах и почти не обнаруживаются в других типах клеток. Это, в том числе, регуляторные петли, необходимые для дифференцировки в миоциты. Механизм формирования белком MyoD петель не установлен.

2.2.4. Активная транскрипция

В недавней работе Friman *at al*. (2022) [26] с помощью данных Hi-C и высокоразрешающего варианта технологии Hi-C под названием micro-C были исследованы взаимодействия между наиболее удалёнными друг от друга последовательностями (линейное расстояние превышало десятки м. п. н.). Сперва в работе рассмотрели, какие белковые факторы расположены на часто контактирующих в пространстве участках. Оказалось, что когезин опосредует взаимодействие на малом расстоянии, белки polycomb — на малом и большом, а транскрипционные факторы — только на большом. Примечательно, что разрушение когезина и удаление polycomb-убиквитинлигазы не приводило к исчезновению дальних взаимодействий, т.е. эти даавно известные архитектурные белки не являются необходимыми для поддержания ультра-длинных петель.

Выяснилось, что далеко взаимодействующие участки являются активными элементами — энхансерами и промоторами, которые преимущественно расположены в *cis* (на одной молекуле ДНК). Таким образом, описанные в статье ULI (ultra-long-range interactions, англ. ультра-дальние взаимодействия) образуют длинные петли размером в десятки м. п. н. с основаниями в промоторах активно транскрибирующихся генов и их энхансерах. Взаимодействующие участки коррелировали с меткой активного хроматина H3K27ac. В клетках разных тканей эта метка находится на разных участках,

так как разные клеточные типы экспрессируют разные гены, и расположение ULI тоже отличается между разными тканями. Также ультра-дальние взаимодействия не зависят от фазы клеточного цикла и восстанавливаются сразу же после митоза. ULI с описанными характеристиками были обнаружены в клетках человека, мыши, *D. rerio* и *D. melanogaster*, то есть формирование очень крупных петель за счёт активной транскрипции встречается как среди позвоночных, так и у беспозвоночных.

У риса (*Oryza sativa*) были обнаружены петли, образованные промотор-промоторными взаимодействиями активных генов. Основания данных петель обогащены РНК-полимеразой II и меткой активного хроматина H3K4me3 (триметилирование четвёртого лизина гистона H3). В этом случае регуляция транскрипции, по всей видимости, является следствием регуляции трёхмерной организации генома [27].

2.2.5. Белки группы polycomb

Белки группы polycomb (PcG) — белки-ремоделлеры хроматина, подавляющие экспрессию генов (репрессоры хроматина). Впервые были обнаружены у *Drosophila melanogaster*, но существуют также у млекопитающих и растений [28]. Получили своё название из-за того, что у мушек, мутантных по генам белков этой группы, появлялось очень много щетинок (англ. *combs*) на задних ногах.

Основная функция белков группы polycomb — подавление активности гомеозисных генов. Гомеозисные (Hox) гены кодируют транскрипционные факторы, которые управляют дифференцировкой тканей и развитием органов в ходе раннего онтогенеза. В конкретной части тела одни группы гомеозисных генов должны функционировать, другие — нет, и белки PcG заключают их в гетерохроматин. Это общая их функция как у дрозофил, так и у млекопитающих. Также PcG участвуют в инактивации X-хромосомы у млекопитающих [29].

У млекопитающих белки PcG подразделяются на две группы — комплексы PRC1 и PRC2. Гены, кодирующие белки PRC1, очень схожи с соответствующими генами плодовых мушек. Комплекс PRC1 распознаёт гистоновую метку H3K27me3 (отмечает гены, подлежащие инактивации) и устанавливает метку H2A-K119ub, которая вызывает компактизацию нуклеосом и превращение участка в гетерохроматин. PRC2 устанавливает на гистоны метки метилирования и этим репрессирует транскрипцию.

Исследования архитектуры генома насекомых на сегодняшний день выполняются в основном только на представителях рода *Drosophila* [30]. ТАДы у плодовых мушек впервые были обнаружены на основе данных Hi-C-эксперимента, выполненного на их эмбрионах. Исследования продемонстрировали, что, в отличие от млекопитающих, у плодовых мушек нет корреляции между местами установки CTCF на ДНК; основания петель обеднены CTCF, хотя на большей их части был обнаружен когезин (его субъединица Rad21). Следовательно, когезин-опосредованная экструзия не является основной причиной петлеобразования [31]. При этом дрозофилы обладают CTCF с последовательностью цинковых пальцев, способной связываться с тем же мотивом, что у млекопитающих. Около 20 % петель расположено в хроматине, репрессированном белками группы Polycomb (Pc), и небольшая их часть — в HP1-репрессированном или активном. При этом почти 40 % оснований всех петель находилось в Pc-репрессированном хроматине и совсем незначительная их часть — в других типах. Выяснилось, что треть всех оснований непосредственно связана с белком Pc, а Pc связывается исключительно с основаниями петель, потому что они содержат “Polycomb response elements” — последовательности, узнаваемые репрессирующим комплексом Polycomb 1 (PRC1); известно, что PCR1 способен компактизировать хроматин *in vitro* [32]. Треть всех этих оснований оказалась промоторами важных для онтогенеза генов — например, Antennapedia (*Antp*) или sex combs reduced (*Scr*). Экспрессия генов под этими промоторами была значительно снижена. Таким образом, сближение комплексом PCR1 конденсированных им участков хроматина формирует петли. Образование петель и регуляция активности генов (а именно — подавление) у *Drosophila* связаны, но причинно-следственные связи пока не установлены. Скорее всего, в этот механизм вовлечены другие белки, так как PRC1 встречается только на трети оснований [33].

2.2.6. Zelda

Транскрипционный и пионерный фактор Zelda (Zld) — белок с цинковыми пальцами, связывается с коротким цис-регуляторным элементом CAGGTAG [34]. Zld отвечает за зиготическую активацию генома (zygotic genome activation, ZGA) у эмбрионов *D. melanogaster*, он «пробуждает» геном [35]. Как и все пионерные факторы, Zld способен узнавать последовательности ДНК, связанные с нуклеосомами и недоступные для обычных транскрипционных факторов. В ходе зиготической активации генома Zld создаёт около 2 000 энхансер-промоторных взаимодействий у

всех эмбриональных клеток независимо от пути дальнейшей дифференцировки, что не согласуется с представлением о клеточной специфичности разных промоторов; возможно, это объясняется дальнейшим устранением части взаимодействий. Эти контакты появляются в ядрах плюрипотентных клеток рано, ещё до начала транскрипции генов и формирования ТАДов, и приводят к образованию хроматиновых петель. Одновременно Zld объединяет далеко отстоящие друг от друга энхансеры в кластеры — *цис*-регуляторные модули (*cis*-regulatory modules, CRMs). Частота контактов энхансеров из одного и того же CRM существенно снижается при деплеции Zld. «Сгущение» этих районов хроматина «разрыхляет» остальные, делая их доступными для других транскрипционных факторов; кроме того, удержание энхансеров в одном кластере препятствует их проникновению в ТАДы, где они не должны быть активны [36]. Ортологи Zld вне пределов класса Членистоногие обнаружены не были.

2.3. Трёхмерная организация генома комаров рода *Anopheles*

В целом пространственная организация хроматина малярийных комаров согласуется со схемой, описанной в разделе «Общие принципы», но существуют важные различия, описанные ниже. Главные особенности 3D-генома *Anopheles* были подтверждены и установлены в работе Lukyanchikova *at al*. (2022) [3] с помощью метода Hi-C.

Размер генома малярийных комаров — порядка 250 м. п. н. [37]. Он представлен тремя линейными хромосомами (2n = 6): половыми (XX у самок, XY у самцов) и двумя субметацентрическими аутосомами (обозначаются номерами: № 2 и № 3) [38]. X-хромосома примерно в три раза меньше аутосом, которые сопоставимы по длине. Для удобства плечи аутосом рассматриваются как отдельные хромосомные элементы (2L с центромерой на конце, 2R с центромерой в начале и так далее). Хромосомы политенизируются в слюнных железах, мальпигиевых сосудах, средней кишке и жировом теле на стадии личинки, а также в мальпигиевых сосудах и питающих клетках яичников на стадии имаго.

По крайней мере в части клеток комара хромосомы организованы по Раблю: все центромеры разных хромосом кластеризованы на одной части ламины, все теломеры — на противоположной, а участки хромосом между ними заполняют пространство ядра, образуя хромосомные территории. Разделение на территории при

этом строгое, межхромосомные контакты у комаров, как и у других насекомых, практически не наблюдаются. В то же время, для многих насекомых и, вероятно, для комаров рода *Anopheles* характерно сближение («спаривание») гомологичных хромосом в интерфазе. При такой организации хромосомы вытянуты продольно, поэтому дальние взаимодействия у них, в целом, отсутствуют. Хромосомные территории имеют эллипсоидную форму, а не сферическую, как у млекопитающих [39]. Интересно, что организация по Раблю была более выраженной у эмбрионов малярийных комаров, нежели у имаго. Конфигурация по Раблю прежде также наблюдалась у дрозофил [40].

Как и хроматин других таксонов, геном малярийных комаров разделён на компартменты активного и неактивного хроматина. На карте контактов Hi-C это отображается наличием «шахматного» рисунка. В отличие от млекопитающих, однако, домены в геномах личинок комаров группируются слабее из-за организации хромосом по Раблю. В свою очередь, компартменты содержат топологически ассоциированные домены (ТАДы) — в эухроматине они меньше и ассоциированы с активной генной экспрессией, в гетерохроматине они крупнее, содержат меньше генов и соответствуют низкому уровню генной экспрессии. Характерный размер ТАДа малярийного комара составляет 200—400 т. п. н.

В статье Lukyanchikova *at al*. (2022) данные, полученные с помощью метода Hi-C, продемонстрировали наличие у пяти видов комаров рода *Anopheles* (*A. albimanus*, *A. atroparvus*, *A. stephensi*, *A. merus*, *A. coluzzii*) взаимодействия между специфичными локусами, которые приводили к образованию петель хроматина. Так как основной механизм образования хроматиновых петель у насекомых — действие белков группы Polycomb, был проведён анализ данных иммунопреципитации хроматина (ChIP-seq) по метке репрессированного хроматина H3K27me3 (триметилирование 27-го лизина гистона H3) в основаниях петель у клеток *A. atroparvus*. Некоторые из них действительно несли данную гистоновую метку, но далеко не все.

Белок CTCF экспрессируется у комаров и способен связываться с теми же последовательностями, что у млекопитающих [41], но столь же значительную роль в организации 3D-генома не играет.

2.4. Длинные хроматиновые петли в геноме комаров рода *Anopheles*

Большинство хроматиновых петель малярийных комаров охватывает геномное расстояние не более чем 1 м. п. н. и располагается в пределах одного ТАДа. Однако в

работе Lukyanchikova *at al*. (2022) были обнаружены петли, простирающиеся на очень большое расстояние — до 31 м. п. н., их количество у каждого из рассмотренных видов составило 2—6. Они включают длинные петли на Х-хромосоме (Х-петли) и на аутосоме (А-петли). Х-петли наблюдались у эмбрионов пяти видов *Anopheles*, линии клеток MSQ43 и взрослых комаров вида *Anopheles merus*. А-петли не наблюдались в линии клеток MSQ43, поэтому эта работа сосредоточена на Х-петлях. Основания длинных петель представляют протяжённые (200—300 т. п. н.) последовательности, которые взаимодействуют значительно чаще, чем ожидается для разделяющего их расстояния. Частота контактов оснований на разных хромосомах не превышает контактность любых других последовательностей на отличных хромосомах, то есть они не колокализуются в пространстве ядра. А-петли и Х-петли (за исключением *A. albimanus*) оказались гомологичны друг другу у всех пяти видов, поскольку в основаниях у них содержатся ортологичные гены.

Далее был осуществлён полногеномный поиск консервативных элементов и определено их распределение в пределах оснований длинных петель. Для установления синтении между видами в качестве референса использовались консервативные элементы *A. atroparvus*. Результаты анализа подтвердили, что петли формируются между гомологичными (синтенными) участками. Часть консервативных элементов обнаруживались у всех пяти видов. Однако у некоторых видов общие консервативные элементы располагались не в тех частях оснований, которые контактировали чаще всего — значит, они не объясняют повышенную частоту контактов.

В пределах некоторых оснований были обнаружены транскрибируемые гены, однако уровень их экспрессии оказался умеренным. Не было также обнаружено обогащения метками активного хроматина. Это наблюдение согласуется с тем, что основания расположены в В-компартменте. Таким образом, формирование длинных Х- и А-петлей не может быть объяснено кластеризацией активного хроматина.

Анализ данных ChIP-seq по метке H3K27me3 (её устанавливают белки группы Polycomb) в клетках *A. atroparvus* показал умеренное обогащение в основаниях Х-петель, но не А-петель. Существуют также блоки хроматина, расположенные между основаниями Х-петель, которые очень сильно обогащены данной меткой, но петель при этом не образуют. Следовательно, взаимодействия, опосредованные белками Polycomb, могут вызывать формирование Х-, но не А-петель. Но и для Х-петель должен существовать дополнительный механизм петлеобразования, поскольку другие участки

генома с таким же уровнем H3K27me3 не демонстрируют такой высокой частоты контактов.

Флуоресцентная гибридизация *in situ* в фолликулярных клетках и питающих клетках яичников комаров видов *A. coluzzii*, *A. stephensi* и *A. atroparvus* показала высокий уровень колокализации проб на основания длинных петель, однако не во всех клетках. В то же время в питающих клетках яичников с высокой степенью политенизации хромосом колокализация проб не обнаружена.

Таким образом, хроматин комаров рода *Anopheles* формирует несколько необычайно длинных петель в определённых участках, которые сохраняются в эволюции более 100 млн лет и формирование которых не может быть объяснено ни одним из известных молекулярных механизмов. В рамках данной работы реализована попытка определить закономерности в строении оснований длинных петель и выдвинуть гипотезу, объясняющую их формирование.

3. Материалы и методы

Вычислительные мощности для работы программного обеспечения, использованного в работе, были предоставлены информационным вычислительным комплексом Новосибирского государственного университета (http://nusc.nsu.ru). Собственное программное обеспечение было разработано на языке Python (версия

3.7.3). Программное обеспечение, использованное в работе, а также параметры запуска программ перечислены в таблице 1.

Таблица 1. Программное обеспечение, использованное в работе.

Задача	**ПО**	**Версия**	**Параметры запуска**
Загрузка данных секвенирования третьего поколения	`wget` (стандартная утилита Linux)	4.12.14-122.136-default	по умолчанию
Выравнивание прочтений на референсный геном	minimap2	2.17	`-t 75`
	bwa mem	0.7.17-r1188	`-t 20`
Конверсия в *.bam, сортировка и индексирование *.sam-файла	view, merge, sort, index (утилиты SAMtools) [42]	1.9	по умолчанию
Визуализация выравнивания	Integrative Genomics Viewer (IGV) [43]	2.13.2	—
Извлечение целевых последовательностей прочтений	собственный Python-скрипт (см. главу 3.1)	—	—
Множественное выравнивание	MAFFT	7.508	`einsi --thread -1 --reorder`
Фильтрация пропусков во множественном выравнивании	trimAl [44]	1.4.rev15	`-fasta -gt 0.25`
Получение консенсусной последовательности	cons (утилита EMBOSS)	6.5.7.0	`-plurality 0 -setcase 0`

Выравнивание консенсусной последовательности на референсную	Unipro UGENE	43.0	—
Аннотация геномных повторов	BuildDatabase	2.0.2	по умолчанию
	RepeatModeler	2.0.2	`-pa 3 -engine=rmblast`
	RepeatMasker	4.1.0	`-pa 3`
Оценка геномного покрытия	bamCoverage (инструмент deepTools) [45]	3.5.1	`-p 3 -of bedgraph`
Извлечение последовательности из FASTA по координатам	getfasta (инструмент BEDtools) [46]	2.26.0	по умолчанию
Поиск гомологичных последовательностей по геномам разных организмов	Nucleotide BLAST (NCBI)	веб-версия	по умолчанию
Поиск химерных сегментов выравнивания по координатам	собственный Python-скрипт (см. главу 3.4)	—	—

3.1. Извлечение прочтений, соответствующих определённым геномным координатам

Для получения последовательностей, соответствующих пустым участкам между ранее собранными контигами (будут обозначаться в тексте как «пробелы», от англ. *gap*), был реализован алгоритм, состоящий из трёх этапов: 0) выравнивание длинных прочтений на геном, пробелы в котором необходимо заполнить; 1) поиск и

уточнение границ пробелов; 2) поиск прочтений, содержащих последовательности пробелов; 3) запись фрагментов выравнивания, соответствующих пробелам.

0. Выравнивание длинных прочтений на референсный геном.

Данные секвенирования третьего поколения — длинные прочтения — выравниваются на референсный геном, содержащий пробелы, которые необходимо заполнить. Полученный в результате файл *.sam конвертируется в файл *.bam, сортируется и индексируется с помощью утилит SAMtools (см. таблицу 1).

1. Поиск и уточнение границ пробелов.

Участки между последовательно идущими контигами в файлах формата *.fasta для доступных нам геномов комаров были во время сборки этих геномов заполнены последовательностями из тысяч подряд идущих букв «N». Определение их границ могло, таким образом, быть формализовано, но на практике в прилежащих к пробелам последовательностях референса при выравнивании наблюдаются неточности, а именно: несоответствия референсному геному (mismatches), аномалии покрытия. Поэтому координаты границ пробелов определяются вручную с помощью визуализации выравнивания в IGV. Установленные границы вносятся в таблицу для автоматического процессирования на следующем шаге работы алгоритма.

2. Поиск прочтений, содержащих последовательности пробелов.

С использованием библиотеки `pysam` языка Python координаты начала и конца сегментов выравнивания, записанных в *.bam-файле, сравниваются с указанными в таблице границами каждого из пробелов. Если сегмент пересекается с пробелом частично или полностью, его последовательность и метаданные сохраняются в памяти.

3. Запись фрагментов выравнивания, соответствующих пробелам.

Далее у каждого из записанных сегментов выравнивания сохраняется, во-первых, участок, соответствующий пробелу в геноме, во-вторых, по обе стороны от этого участка — последовательности длиной 1,5 т. п. н. (фланги). Эти три части записаны как одна последовательность в формате FASTA. Остальная часть сегмента обрезается.

Фланкирующие участки необходимы для того, чтобы впоследствии, совершив множественное выравнивание и получив консенсус, выровнять его на референсный геном и точно позиционировать участок между ними, который соответствует пробелу. Это необходимо, потому что заранее не известно, какой именно длины должна быть

последовательность на месте пробела (количество букв «N» при сборке было выбрано произвольно).

Описанный алгоритм можно применять не только для заполнения пробелов в сборках геномов, но в целом для извлечения из *.bam-файла прочтений, накладывающихся на область с известными координатами. К примеру, если при выравнивании выясняется, что прочтения содержат инсерцию, алгоритм может получить участки с инсерцией из множества прочтений, которые затем можно множественно выровнять и получить точный консенсус вставки.

3.2. Описание геномных регионов *A. stephensi*, использованных в работе

Таблица 2. Описание геномных регионов *A. stephensi*, использованных в работе.

Задача	**Регион**	**Хромосома**	**Координата начала региона**	**Координата конца региона**
Извлечение целевых последовательностей прочтений, аннотация геномных повторов	«левое» основание X-петли	X	9 440 002	9 720 570
	«правое» основание X-петли	X	14 660 541	15 057 855
	Центромерный регион	X	16 190 858	19 770 000 (конец хромосомы)

3.3. Аннотация геномных повторов в основаниях X-петель

Для анализа повторённых последовательностей генома были применены программы BuildDatabase, RepeatModeler и RepeatMasker.

BuildDatabase принимает геномный файл FASTA (с заполненными на предыдущих этапах пробелами) и, выравнивая геном сам на себя, создаёт базу данных повторённых элементов. RepeatModeler структурирует эту базу и классифицирует элементы. Наконец, RepeatMasker создаёт таблицу, в которой перечислены все повторённые элементы, указаны их положения в геноме и распределение по классам.

Работа с результирующей таблицей велась с помощью программной библиотеки `pandas` на языке Python.

3.4. Поиск химерных сегментов выравнивания по координатам

Пусть последовательность X записана в референсном геноме один раз, но на самом деле (в реальных молекулах ДНК) повторяется чаще. Тогда некоторые прочтения генома, содержащие X, будут выравниваться химерно: фрагмент с X выровняется на однократно записанный участок, а остальная часть прочтения — с разрывом на другой локус (см. рисунок 1). Необходимо определить, где на самом деле локализованы остальные копии X.

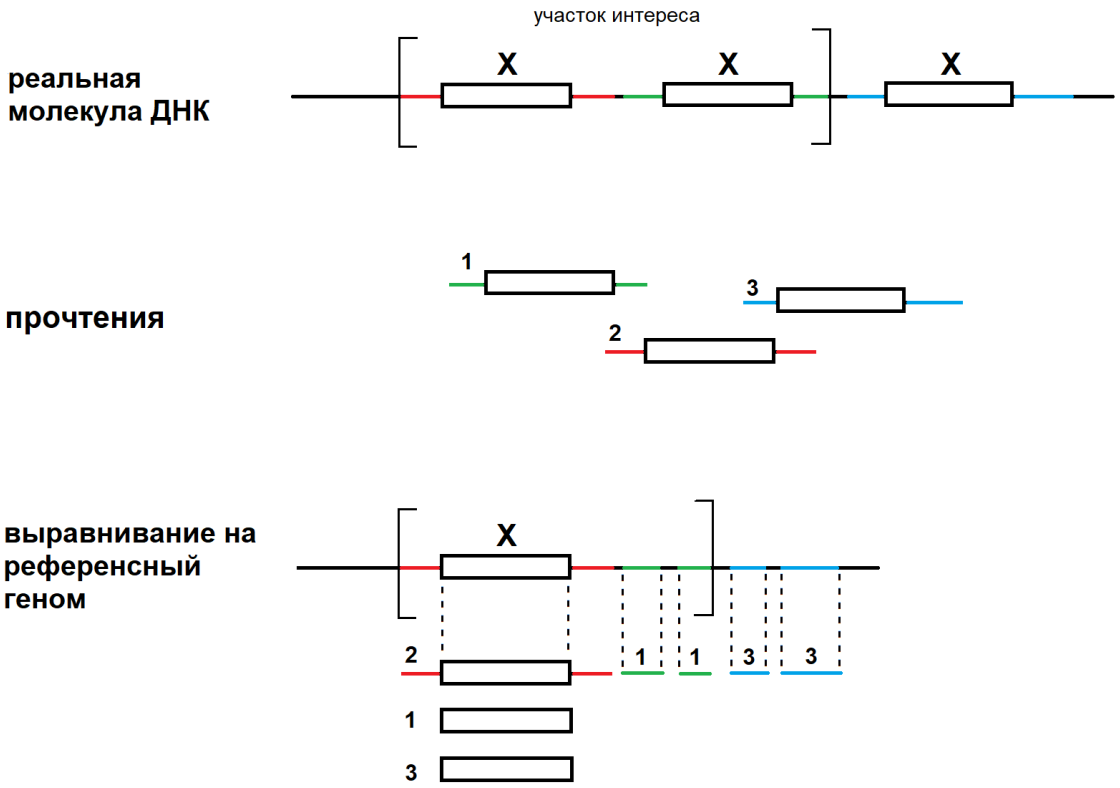

Рисунок 1. Образование химерных выравниваний. Числами обозначены прочтения и их сегменты выравнивания.

Для выполнения этой задачи был реализован следующий алгоритм:

1. Выровнять прочтения на референсный геном и получить *.bam-файл выравнивания.

2. Используя *.bam-файл, сохранить названия всех прочтений, которые выравниваются на последовательность X в референсе (необходимо указать координаты начала и конца X).
3. Выполнить поиск сегментов выравнивания прочтений, названия которых были сохранены на предыдущем этапе, в участке интереса (указав его координаты), исключая референсную последовательность X.

Пункты 2 и 3 алгоритма реализованы с использованием библиотеки `pysam` языка Python.

Таким образом, обнаружение дополнительных сегментов выравнивания в пункте 3 алгоритма будет основанием подозревать повторение последовательности X в участке интереса. Прямое доказательство можно получить, изучив все сегменты подозрительного прочтения в *.bam-файле с помощью стандартной утилиты Linux `grep`.

4. Результаты и обсуждение

4.1. Загрузка данных секвенирования третьего поколения

На сегодняшний день сборки геномов малярийных комаров несовершенны. В районе оснований длинных петель контиги разделены пустыми последовательностями-пробелами (заполнены буквами «N») суммарной длины около нескольких десятков т. п. н. Предполагается, что пробелы образовались из-за невозможности при секвенировании по технологии Illumina однозначно локализовать

участки генома с повторёнными элементами. Мы приняли решение использовать данные секвенирования третьего поколения, которые отличаются большой длиной прочтений (десятки т. п. н.), чтобы заполнить пробелы и в дальнейшем работать с дополненными сборками геномов.

Прочтения геномов пяти видов малярийных комаров, полученные с применением технологий Oxford Nanopore и Pacific Biosciences SMRT, были загружены из базы данных NCBI Sequence Read Archive (SRA) в хранилище вычислительного комплекса НГУ. Список источников см. в приложении 1. На текущем этапе работы мы сосредоточились на одном виде (*A. stephensi*), поскольку в нашей лаборатории есть модельная клеточная линия, полученная из личинок комаров этого вида, а также поскольку для *A. stephensi* в публичных репозиториях доступно большое количество данных секвенирования третьего поколения. В геноме *A. stephensi* мы сфокусировались на X-петле, поскольку, как было показано ранее, эта петля формируется по особому механизму и является эволюционно консервативной. Таким образом, перед нами стояла задача заполнить пробелы в основаниях X-петли *A. stephensi* на основе публично доступных данных секвенирования третьего поколения.

4.2. Выравнивание и визуализация прочтений

Загруженные прочтения генома *Anopheles stephensi* были выровнены на геномную сборку AsteI2_V4 (https://genedev.bionet.nsc.ru/Anopheles.html). Мы выбрали алгоритм minimap2, так как при работе с длинными прочтениями (длиной более 1 т. п. н.) он превосходит традиционный алгоритм Burrows–Wheeler Aligner в точности и скорости выравнивания [47]. Выравнивание и покрытие генома прочтениями были визуализированы для контроля качества данных секвенирования (см. рисунок 2).

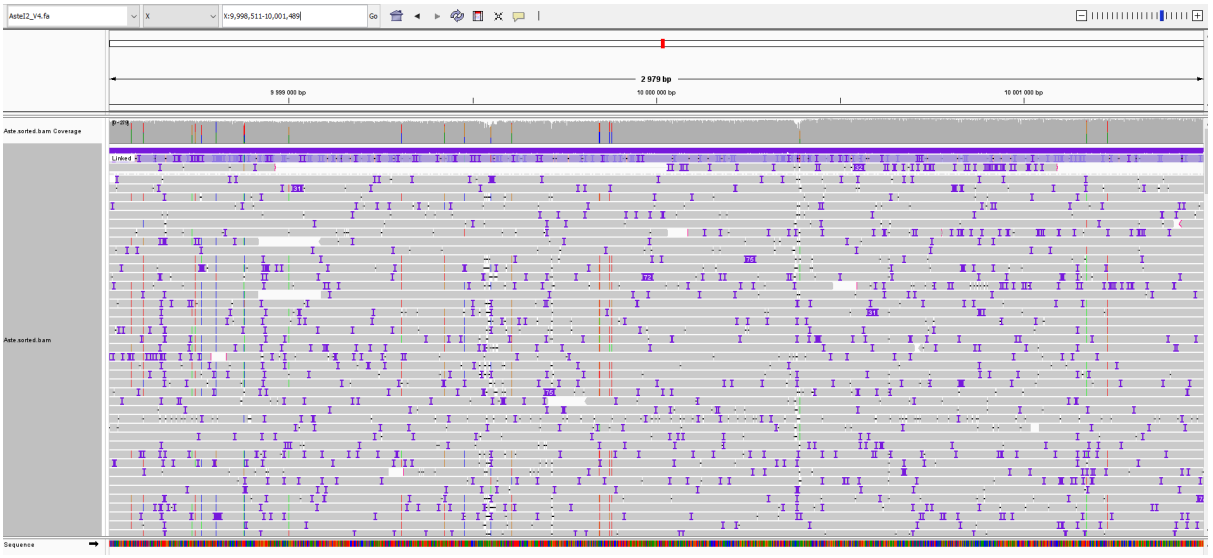

Рисунок 2. Визуализация выравнивания длинных прочтений на референсный геном в программе Integrative Genomics Viewer (IGV).

4.3. Извлечение прочтений, соответствующих определённым геномным координатам

Мы применили алгоритм, описанный в главе 3.1, для извлечения из файла с выравниваниями прочтений, приходящихся на пробелы геномной сборки в основаниях длинных X-петель. Первоначально мы получали около 700 последовательностей за запуск. Но дальнейшее множественное выравнивание такого большого количества участков происходило очень медленно. Поэтому максимальное количество сохраняемых последовательностей было решено снизить до 200 — это ускоряет работу выравнивателя и не снижает качество консенсуса (см. подробнее таблицу 3 в главе 4.5). Таким образом, на основе результатов выравнивания длинных прочтений нами были получены последовательности, содержащие ранее не заполненные районы в основаниях X-петли.

4.4. Множественное выравнивание и получение консенсуса

Секвенирование третьего поколения позволяет прочитывать длинные последовательности, однако является неточным — в полученных последовательностях часто встречаются ошибки [48]. Чтобы устранить ошибки, было решено построить консенсусную последовательность для каждого из исследуемых районов X-петли. Для этого мы провели множественное выравнивание извлечённых в формате FASTA последовательностей. Для работы выравнивателя мы выбрали режим E-INS-i,

поскольку он учитывает наличие невыравниваемых участков между выравниваемыми [49]. На вывод программа производит FASTA-файл множественного выравнивания, который может быть представлен в виде таблицы: все строки в нём одинаковой длины и составлены из пяти символов — четырёх нуклеотидов (A, T, G, C) и дефиса (-). Колонка таблицы — это символы, содержащиеся на одинаковых позициях в каждой из строк. Выравниватель стремится разместить символы в строках так, чтобы в каждой колонке преобладал какой-то один из пяти. Он не может менять символы местами, но может вставлять дефисы. Нуклеотид, который повторяется в колонке чаще всех остальных, будет признан консенсусным. Длина консенсусной последовательности, таким образом, будет равна числу колонок и содержать только консенсусные нуклеотиды.

Как было указано выше, прочтения, полученные методами секвенирования третьего поколения, содержат большое количество ошибок, в том числе инсерций, то есть вставок произвольного количества символов. Поскольку такие ошибки уникальны для каждого прочтения, при построении таблицы, описанной выше, сформируется много колонок, по большей части состоящих из дефисов. Небольшая доля нуклеотидов в этих колонках — ошибки секвенирования. Тем не менее, на этих позициях они будут входить в консенсус. Чтобы этого не происходило, было решено устранять колонки с большим количеством дефисов. Мы выбрали 25-процентный порог: если в колонке более 75 % всех символов — дефисы, она будет удалена из таблицы. Для такой фильтрации мы использовали инструмент trimAl. После фильтрации из файла множественного выравнивания была извлечена консенсусная последовательность.

4.5. Интеграция консенсусных последовательностей в сборку генома

После получения консенсусных последовательностей пробелов, содержащихся в основаниях петель, перед нами встала задача добавить полученные консенсусы в основную сборку генома *A. stephensi*. При извлечении из фрагментов выравнивания последовательностей, соответствующих пробелам в основаниях петель, были сохранены фланкирующие участки (см. главу 3.1). После извлечения консенсуса эти участки были использованы для выравнивания на референсный геном в программе UGENE (см. рисунок 3).

Рисунок 3. Выравнивание извлечённой из ридов референсной последовательности («1» на рисунке) на референсный геном («2» на рисунке). Красными стрелками обозначены границы между пробелом и фланкирующими частями.

По визуализации данного выравнивания можно определить, какая часть консенсуса принадлежит геному, а какая часть должна быть расположена на месте пробела. После этого необходимая последовательность вручную записывается вместо пробела (координаты которого определялись на первом шаге алгоритма, описанного в главе 3.1).

Затем исходные прочтения (глава 4.1) были снова выровнены, но уже на дополненную сборку генома. Выравнивание было визуализировано в IGV для контроля качества сборки (см. рисунок 4).

Рисунок 4. Выравнивание длинных прочтений на референсный геном до и после заполнения пробелов. Визуализация в IGV.

Как было упомянуто ранее, множественное выравнивание нескольких сотен последовательностей занимает очень много времени и вычислительных ресурсов. Поэтому мы решили проверить, с какой скоростью будут множественно выравниваться 2, 5, 10, 50 и 100 фланкированных последовательностей и как будут выравниваться на референс полученные из них консенсусы. Результаты теста представлены на таблице 3:

Таблица 3. Зависимость скорости множественного выравнивания от количества последовательностей.

Количество ридов в *.fa-файле	**Время множественного выравнивания, сек**
2	2
5	40

10	80
50	1711
100	2160

Для оптимизации скорости множественного выравнивания была также отрегулирована длина фланкирующих последовательностей (в итоге она, как было указано выше, составляет 1,5 т. п. н.).

В результате теста выяснилось, что консенсус, полученный даже из десяти фланкированных последовательностей, хорошо выравнивается на референсный геном (см. рисунок 5).

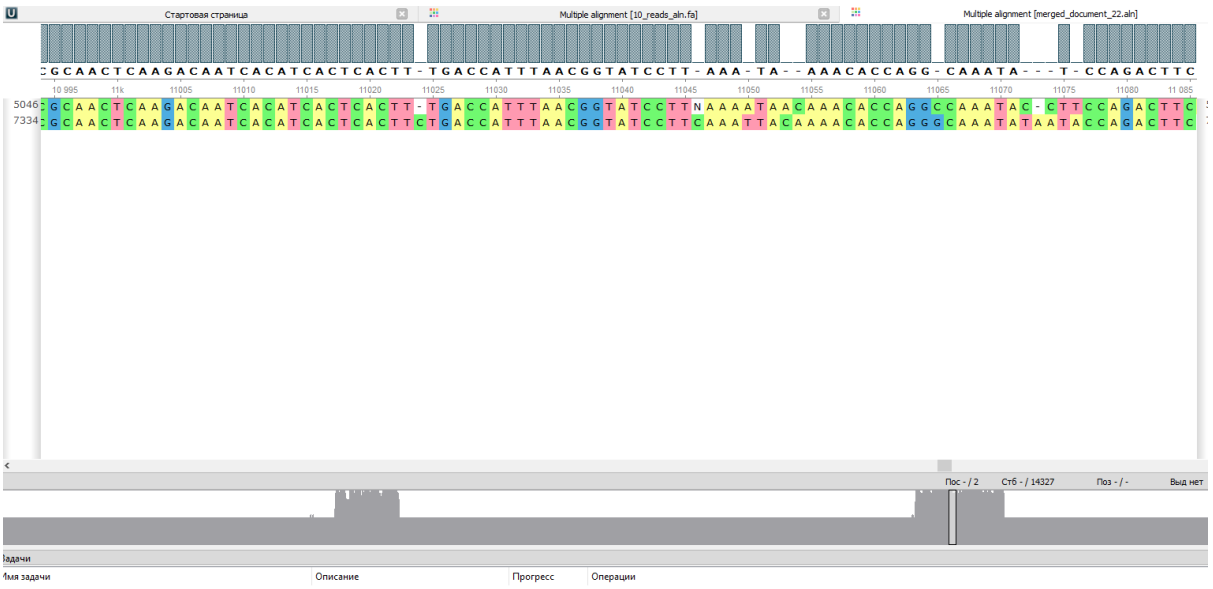

Рисунок 5. Выравнивание флангов консенсуса, полученного множественным выравниванием 10 последовательностей, на референсный геном. Визуализация в Unipro UGENE.

После перечисленных шагов мы получили версию генома *A. stephensi* с заполненными пробелами в основаниях длинных петель, которую можно применять в дальнейшем исследовании.

4.6. Аннотация геномных повторов в основаниях X-петель

Основания X-петель удалены друг от друга на огромное расстояние (до десятков м. п. н.), но в то же время контактируют с очень большой частотой. Можно предположить, что столь мощное взаимодействие должно быть обусловлено большим количеством белков. Многократное повторение сайта связывания позволило бы большому числу молекул белка, аффинного к данному сайту, связаться с основанием петли, что позволило бы объяснить наблюдаемое взаимодействие.

В дополнение к этому, основания X-петель находятся в B-компартменте, то есть в гетерохроматине. Известно, что конститутивный гетерохроматин содержит большое количество повторённых последовательностей — предположительно, конденсация хроматина является одним из механизмов сдерживания экспрессии повторённых элементов [50].

Из этих соображений мы приняли решение проанализировать повторы в дополненных последовательностях оснований X-петель. Мы сравнили обогащение повторами (как отношение общей длины всех повторов к длине анализируемого участка) оснований длинных петель и остальной части хромосомы, а также оценили количество повторов в центромерной области в качестве контроля (координаты оснований петель и центромерного региона см. в таблице 2, глава 3.2). Повторы составляют 10,8 % X-хромосомы без учёта прицентромерной части (и без центромеры, и без оснований петель — 10,7 %). В основаниях петель повторы занимают 12,4 % ДНК, в центромере — 25 %. Как и ожидалось, центромера богаче повторёнными элементами. Таким образом, основания длинных петель содержат больше повторов, чем вся хромосома в среднем. В дальнейшем мы планируем статистически проверить значимость этого различия.

Мы предположили, что основания могут быть обогащены только одним (или несколькими) конкретными классами повторов, и проанализировали покрытие тех же самых областей, но уже повторами отдельных классов. Для этого при помощи программы RepeatModeler мы разделили все выявленные повторы на 15 классов. Полученные значения представленности каждого из классов повторов указаны в приложении 3. Основания длинных петель оказались несколько богаче LINE-элементами, чем остальная хромосома. В дальнейшем мы планируем более внимательно изучить распределение этих элементов.

Наконец, классы повторов содержат подклассы, и основания длинных петель могут значительно отличаться от остальной хромосомы по их содержанию. Наиболее часто встречающиеся в основаниях петель подклассы повторов приведены в приложении 4. Действительно, есть подклассы, которые встречаются в основаниях до 24 раз чаще, чем на остальной X-хромосоме. Они представляют повторы коротких последовательностей, 4—9 п. н. Заслуживает внимания факт, что в первой десятке каждый подкласс в избытке встречается на одном основании и полностью отсутствует на другом. Ожидалось, что повторы, ответственные за пространственный контакт, будут распределены одинаково на обоих основаниях. Поэтому мы считаем, что обогащение данными подклассами повторов не связано с формированием X-петли.

4.7. Изучение областей генома с аномальным покрытием

Мы уделили особое внимание участкам генома с аномально высоким покрытием при выравнивании длинных прочтений (на несколько порядков выше, чем в среднем по геному). Если некая последовательность вошла в сборку генома в единственном экземпляре, но в реальных молекулах ДНК повторяется значительно чаще, при выравнивании её покрытие будет выше, чем в среднем по геному. Таким образом, эта последовательность не будет распознаваться как повторённая и ускользнёт от программ-анализаторов.

Мы оценили покрытие оснований длинных петель. В левом основании длинной петли был обнаружен участок, покрытие которого более чем в 300 раз превышает среднее по геному (110 000 против 300, см. рисунок 6). Поиск гомологии показал, что найденный участок является фрагментом предсказанной последовательности рРНК большой субъединицы рибосомы. Казалось странным, что в регионе, который находится в B-компартменте, может располагаться ген, для которого характерна активная экспрессия. К тому же, найденная нами последовательность составляла только фрагмент гена.

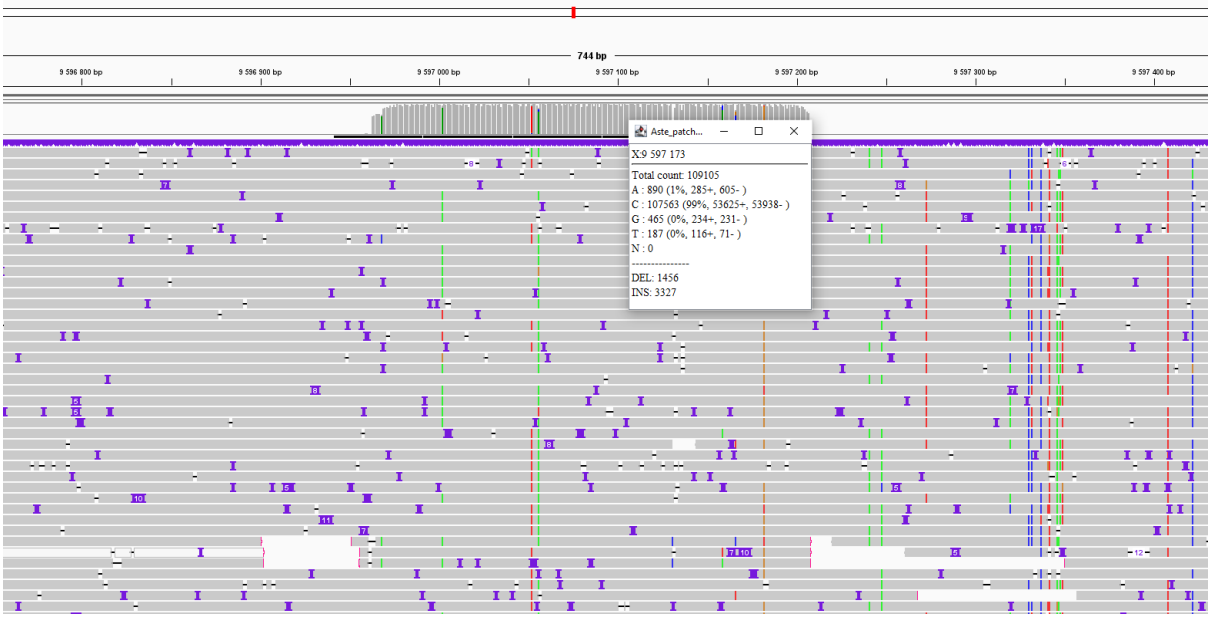

Рисунок 6. Обнаруженный участок с аномальным покрытием. Визуализация в IGV.

С другой стороны, наличие генов рРНК в основаниях длинных петель могло бы объяснить их формирование — в таком случае основания являлись бы ядрышковыми организаторами на X-хромосомах *A. stephensi*. Поэтому мы решили найти в NCBI все последовательности генов рибосомных РНК *A. stephensi* и выровнять их на геном. В итоге гены рРНК выравнивались не на основания длинных петель и даже не на X-хромосому. Единственный сегмент выравнивания, пришедшийся на основания, оказался уже обнаруженным нами участком. Поэтому мы пришли к выводу, что основания длинных X-петель не могут являться ядрышковыми организаторами.

Как было упомянуто ранее, высокое покрытие участка может свидетельствовать о том, что в реальном геноме он повторён многократно, а в референсе записан однократно. Поэтому прочтения, которые содержат повтор участка из другой области генома, будут выравниваться на единственное его вхождение в референсе и давать большое покрытие. Нас интересовало, повторяется ли обнаруженный участок в левом основании длинной петли. Чтобы понять это, мы применили алгоритм, описанный в главе 3.4. Многие сегменты выравнивания, которые выровнялись на сверхпокрытый участок, принадлежат прочтениям, основное выравнивание которых приходится на другую хромосому. Также они имеют множество дополнительных сегментов выравнивания в прицентромерном районе X-хромосомы. Мы не обнаружили прочтений, включающих участок рибосомального гена и участок основания левой

петли, не прилежащий непосредственно к рибосомальному гену. Из этого мы сделали вывод, что сверхпокрытый участок расположен в левом основании однократно.

5. Предварительные выводы

Использование данных секвенирования третьего поколения позволило заполнить пробелы геномных последовательностях оснований длинных петель на хромосоме X комаров *A. stephensi*

В основаниях длинной хроматиновой X-петли комара *Anopheles stephensi* были выявлены и охарактеризованы повторённые элементы. Наблюдается различие в обогащении LINE-элементами оснований петель и остальной части хромосомы.

В одном из оснований длинной X-петли комара *Anopheles stephensi* был обнаружен ген рРНК, однако он не повторяется в обоих основаниях и, скорее всего, является псевдогеном, который случайно расположился в обнаруженной локации.

В будущем планируется:

1. Усовершенствовать сборки геномов ещё четырёх видов малярийных комаров (*A. albimanus*, *A. atroparvus*, *A. merus* и *A. coluzzii*).
2. Провести более детальный статистический анализ распределения повторённых элементов на X-хромосоме *Anopheles stephensi* (в том числе с использованием метода bootstrap). Выявить, охарактеризовать и статистически проанализировать повторы у четырёх других видов комаров.
3. Провести межвидовое сравнение распределения повторённых последовательностей у пяти видов малярийных комаров.

Список цитируемой литературы

1. Ghavi-Helm Y, Jankowski A, Meiers S, Viales RR, Korbel JO, Furlong EEM. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nat Genet. 2019;51:1272–82.
2. Renschler G, Richard G, Valsecchi CIK, Toscano S, Arrigoni L, Ramírez F, et al. Hi-C guided assemblies reveal conserved regulatory topologies on X and autosomes despite extensive genome shuffling. Genes Dev. 2019;33:1591–612.
3. Lukyanchikova V, Nuriddinov M, Belokopytova P, Taskina A, Liang J, Reijnders MJMF, et al. Anopheles mosquitoes reveal new principles of 3D genome organization in insects. Nat Commun 2022 131. 2022;13:1–22.
4. Marchal C, Sima J, Gilbert DM. Control of DNA replication timing in the 3D genome. Nat Rev Mol Cell Biol 2019 2012. 2019;20:721–37.
5. van Steensel B, Furlong EEM. The role of transcription in shaping the spatial organization of the genome. Nat Rev Mol Cell Biol 2019 206. 2019;20:327–37.
6. Piazza A, Bordelet H, Dumont A, Thierry A, Savocco J, Girard F, et al. Cohesin regulates homology search during recombinational DNA repair. Nat Cell Biol 2021 2311. 2021;23:1176–86.
7. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nat Rev Genet 2018 1912. 2018;19:789–800.
8. Kim A, Dean A. Chromatin loop formation in the β-globin locus and its role in globin gene transcription. Mol Cells. 2012;34:1.
9. Zhang Y, Zhang X, Ba Z, Liang Z, Dring EW, Hu H, et al. The fundamental role of chromatin loop extrusion in physiological V(D)J recombination. Nat 2019 5737775. 2019;573:600–4.
10. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science. 2009;326:289.
11. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 2015;10:1297–309.
12. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature.

2012;485:376–80.
13. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.
14. Yusufzai TM, Tagami H, Nakatani Y, Felsenfeld G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. Mol Cell. 2004;13:291–8.
15. Sanborn AL, Rao SSP, Huang SC, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A. 2015;112:E6456–65.
16. Nora EP, Goloborodko A, Valton AL, Gibcus JH, Uebersohn A, Abdennur N, et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. Cell. 2017;169:930-944.e22.
17. Haarhuis JHI, van der Weide RH, Blomen VA, Yáñez-Cuna JO, Amendola M, van Ruiten MS, et al. The cohesin release factor WAPL restricts chromatin loop extension. Cell. 2017;169:693-707.e14.
18. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nat 2013 5027469. 2013;502:59–64.
19. Ganji M, Shaltiel IA, Bisht S, Kim E, Kalichava A, Haering CH, et al. Real-time imaging of DNA loop extrusion by condensin. Science. 2018;360:102–5.
20. Bintu B, Mateo LJ, Su JH, Sinnott-Armstrong NA, Parker M, Kinrot S, et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science. 2018;362.
21. Beagan JA, Duong MT, Titus KR, Zhou L, Cao Z, Ma J, et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res. 2017;27:1139–52.
22. Weintraub AS, Li CH, Zamudio A V., Sigova AA, Hannett NM, Day DS, et al. YY1 is a structural regulator of enhancer-promoter loops. Cell. 2017;171:1573-1588.e28.
23. Faerman A, Goldhamer DJ, Puzis R, Emerson CP, Shani M. The distal human myoD enhancer sequences direct unique muscle-specific patterns of lacZ expression during mouse development. Dev Biol. 1995;171:27–38.
24. Wang R, Chen F, Chen Q, Wan X, Shi M, Chen AK, et al. MyoD is a 3D genome structure organizer for muscle cell identity. Nat Commun 2022 131. 2022;13:1–17.

25. Conerly ML, Yao Z, Zhong JW, Groudine M, Tapscott SJ. Distinct activities of Myf5 and MyoD indicate separate roles in skeletal muscle lineage specification and differentiation. Dev Cell. 2016;36:375–85.
26. Friman ET, Flyamer IM, Boyle S, Bickmore WA. Ultra-long-range interactions between active regulatory elements. bioRxiv. 2022;:2022.11.30.518557.
27. Zhao L, Wang S, Cao Z, Ouyang W, Zhang Q, Xie L, et al. Chromatin loops associated with active genes and heterochromatin shape rice genome architecture for transcriptional regulation. Nat Commun 2019 101. 2019;10:1–13.
28. Jiao H, Xie Y, Li Z. Current understanding of plant polycomb group proteins and the repressive histone H3 lysine 27 trimethylation. Biochem Soc Trans. 2020;48:1697–706.
29. Blackledge NP, Klose RJ. The molecular principles of gene regulation by polycomb repressive complexes. Nat Rev Mol Cell Biol 2021 2212. 2021;22:815–33.
30. Liao Y, Zhang X, Chakraborty M, Emerson JJ. Topologically associating domains and their role in the evolution of genome structure and function in Drosophila. Genome Res. 2021;31:397–410.
31. Gambetta MC, Furlong EEM. The insulator protein CTCF is required for correct Hox gene expression, but not for embryonic development in Drosophila. Genetics. 2018;210:129–36.
32. Francis NJ, Kingston RE, Woodcock CL. Chromatin compaction by a polycomb group protein complex. Science. 2004;306:1574–7.
33. Eagen KP, Aiden EL, Kornberg RD. Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. Proc Natl Acad Sci U S A. 2017;114:8764–9.
34. Ventos-Alfonso A, Ylla G, Belles X. Zelda and the maternal-to-zygotic transition in cockroaches. FEBS J. 2019;286:3206–21.
35. Liang HL, Nien CY, Liu HY, Metzstein MM, Kirov N, Rushlow C. The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. Nat 2008 4567220. 2008;456:400–3.
36. Hug CB, Grimaldi AG, Kruse K, Vaquerizas JM. Chromatin architecture emerges during zygotic genome activation independent of transcription. Cell. 2017;169:216-228.e19.
37. Zamyatin A, Avdeyev P, Liang J, Sharma A, Chen C, Lukyanchikova V, et al. Chromosome-level genome assemblies of the malaria vectors Anopheles coluzzii and Anopheles arabiensis. Gigascience. 2021;10.
38. Artemov GN, Stegniy VN, Sharakhova M V., Sharakhov I V. The development of

cytogenetic maps for malaria mosquitoes. Insects. 2018;9.
39. George P, Kinney NA, Liang J, Onufriev A V., Sharakhov I V. Three-dimensional organization of polytene chromosomes in somatic and germline tissues of malaria mosquitoes. Cells. 2020;9.
40. Moretti C, Stévant I, Ghavi-Helm Y. 3D genome organisation in Drosophila. Brief Funct Genomics. 2020;19:92–100.
41. Gray CE, Coates CJ. Cloning and characterization of cDNAs encoding putative CTCFs in the mosquitoes, Aedes aegypti and Anopheles gambiae. BMC Mol Biol. 2005;6.
42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
43. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol 2011 291. 2011;29:24–6.
44. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972.
45. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42:W187–91.
46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
47. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
48. Delahaye C, Nicolas J. Sequencing DNA with nanopores: troubles and biases. PLoS One. 2021;16.
49. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
50. Janssen A, Colmenares SU, Karpen GH. Heterochromatin: guardian of the genome. https://doi.org/101146/annurev-cellbio-100617-062653. 2018;34:265–88.

Приложения

Приложение 1. Источники прочтений секвенирования третьего поколения.

Вид	**Страница в архиве SRA**	**Технология**
A. stephensi	https://www.ncbi.nlm.nih.gov/sra/?term=SRR1168957	PacBio
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR11672505	
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR11672506	
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR15605613	
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR17013320	
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR17013321	
A. albimanus	https://www.ncbi.nlm.nih.gov/sra/?term=SRR11570338	Oxford Nanopore
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR17873073	PacBio SMRT
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR17873074	
A. atroparvus	https://www.ncbi.nlm.nih.gov/sra/?term=ERR3610908	PacBio SMRT
	https://www.ncbi.nlm.nih.gov/sra/?term=ERR6511325	
	https://www.ncbi.nlm.nih.gov/sra/?term=ERR6511326	
A. coluzzii	https://www.ncbi.nlm.nih.gov/sra/?term=SRR11915740	Oxford Nanopore
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR13040395	
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR13040396	
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR13040398	
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR13040401	
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR13040402	
	https://www.ncbi.nlm.nih.gov/sra/?term=SRR13040397	PacBio SMRT
A. merus	https://www.ncbi.nlm.nih.gov/sra/?term=SRR17020779	PacBio SMRT

Приложение 2. Координаты пробелов в основаниях длинных петель *A. stephensi* (сборка генома AsteI2_V4).

Элемент	**Координата начала**	**Координата конца**	**Элемент**	**Координата начала**	**Координата конца**
Левое основание длинной петли	9435000	9815000	Правое основание длинной петли	14665000	15130000
Пробел № 1.1	9459201	9463274	Пробел № 2.1	14790135	14794474
Пробел № 1.2	9479621	9481049	Пробел № 2.2	14799226	14799911
Пробел № 1.3	9546122	9549394	Пробел № 2.3	14849914	14851407
Пробел № 1.4	9550663	9552170	Пробел № 2.4	14855734	14858902
Пробел № 1.5	9554736	9559366	Пробел № 2.5	14928794	14931978
Пробел № 1.6	9566069	9567074	Пробел № 2.6	14932694	14938656
Пробел № 1.7	9580313	9588396	Пробел № 2.7	14940441	14942842
Пробел № 1.8	9597978	9601003	Пробел № 2.8	14947360	14951120
Пробел № 1.9	9606503	9607807	Пробел № 2.9	14994288	14994731
Пробел № 1.10	9649860	9654830			
Пробел № 1.11	9657562	9659081			
Пробел № 1.12	9662511	9663344			
Пробел № 1.13	9675550	9680512			
Пробел № 1.14	9698331	9699752			

Приложение 3. Покрытие оснований длинных петель *A. stephensi* разными классами повторённых последовательностей (все значения указаны в % × 10^{-3} длины нуклеотидной последовательности).

Класс повтора	**X-хромосома за исключением центромеры**	**X-хромосома за исключением центромеры и оснований длинной петли**	**Левое основание длинной петли**	**Правое основание длинной петли**	**Оба основания длинной петли**	**Центромера**
DNA/TcMar-Mariner	21	22	0	0	0	64
DNA/TcMar-Tc1	109	108	36	229	149	203
DNA/hAT-Tip100	87	84	391	0	162	12
LINE/CR1	307	301	742	197	423	804
LINE/I	329	324	457	439	446	5640
LINE/I-Jockey	71	43	1305	265	695	1152
LINE/R1	85	79	389	106	224	1073
LINE/RTE	1171	1184	1145	670	867	547
LINE/RTE-BovB	850	839	1205	1033	1104	605
LTR/Gypsy	9	9	0	0	0	1267
LTR/Pao	7	7	0	0	0	576
Low complexity	366	368	270	354	319	125
Satellite	22	22	0	24	14	8
Simple repeat	2195	2200	1816	2272	2083	756
Unknown	5063	4991	7599	6087	6713	8298

Приложение 4. Покрытие оснований длинных петель *A. stephensi* разными подклассами повторённых последовательностей (все значения указаны в % × 10^{-5} длины нуклеотидной последовательности).

Подкласс повтора	**X-хромосома за исключением центромеры**	**Левое основание длинной петли**	**Правое основание длинной петли**	**Оба основания длинной петли ÷ X-хромосома за исключением центромеры**	**Центромера**
(GGATGA)n	21	0	856	23,90476	0
(CACATA)n	34	0	1385	23,88235	113
(TCCCGC)n	170	0	6922	23,86471	0
(TTCCTC)n	52	0	2115	23,84615	0
(CGAAGC)n	54	0	2190	23,77778	0
(AGCTCC)n	23	0	932	23,73913	0
(CATCT)n	38	0	1536	23,68421	0
(AGATC)n	20	1141	0	23,65	0
(ACTG)n	30	1711	0	23,63333	0
(GATAGA)n	30	0	1209	23,63333	0
(GGTGGTG)n	30	0	1209	23,63333	0
(TGCTAT)n	32	1818	0	23,53125	0
(TCATTT)n	27	0	1083	23,51852	0
(ATCTTAT)n	22	0	881	23,5	0
(GTAGAA)n	22	0	881	23,5	0
(TAGAGA)n	22	1248	0	23,5	0
(CCGTAT)n	39	0	1561	23,46154	0
(TGGAGC)n	29	0	1158	23,41379	0
(GTGCAGT)n	24	0	957	23,375	0
(ATATG)n	36	963	756	23,36111	0