Московский государственный университет имени М.В.Ломоносова

Биологический факультет

Кафедра биоинженерии

# Разработка подходов интегративного моделирования структуры и динамики нуклеосомных фибрилл на основе анализа экспериментальных данных 3D-геномики субнуклеосомного разрешения

Выпускная квалификационная

работа бакалавра

студента IV курса

**ТИМОХИНА Григория Сергеевича**

Научный руководитель

д.ф.-м.н.

**Шайтан Алексей Константинович**

Москва

2021

СПИСОК СОКРАЩЕНИЙ

п.о. -- пара оснований

ДНК -- дезоксирибонуклеиновая кислота

ОГЛАВЛЕНИЕ

# 1. **ВВЕДЕНИЕ**

ДНК в ядрах клеток эукариот компактизована примерно в $10^5$ раз, образуя вместе со специальными белками сложно организованный нуклеопротеидный комплекс — хроматин. При этом ДНК активно транскрибируется, причем регуляция транскрипции во многом опосредуется структурой хроматина и ее динамическими перестройками. Изучение принципов и закономерностей пространственной организации генома и ее связи с регуляцией транскрипции является одной из основных задач современной молекулярной биологии[1]. Выделяют несколько уровней компактизации хроматина[2]. Высший уровень — уровень хромосом в метафазном хроматине и соответствующий ему уровень хромосомных территорий в интерфазном хроматине — детектируется с помощью оптической микроскопии. Низший уровень — уровень нуклеосомной фибриллы — был открыт, благодаря электронной микроскопии. Нуклеосомная фибрилла представляет собой ряд связанных с ДНК гистоновых октамеров -- нуклеосом -- соединенных линкерными участками ДНК[3]. Срединные уровни были изучены с использованием метода Hi-C, позволяющего определять частоту пространственных взаимодействий между всеми участками генома заданной длины[4]. Несмотря на широкие возможности метода Hi-C и методов современной электронной микроскопии, их разрешающая способность не позволяет исследовать супрануклеосомный уровень компактизации хроматина — уровень компактизации нуклеосомной фибриллы, на котором происходит сближение промоторов и соответствующих им цис-регуляторных элементов. При этом, по мнению многих исследователей, именно на супрануклеосомном уровне происходит регуляция транскрипции, опосредованная динамическими перестройками структуры хроматина[5]. Для изучения супрануклеосомного уровня компактизации хроматина было разработано множество подходов

молекулярного моделирования структуры и динамики нуклеосомной фибриллы[6]. Однако результаты такого моделирования зависят от заданных перед в начале параметров, поэтому для получения реалистичных результатов необходимо основывать моделирование на экспериментальных данных. Таким образом перспективными для изучения супрануклеосомного уровня компактизации являются интегративные подходы, то есть подходы, сочетающие моделирование и анализ экспериментальных данных. Важным параметром для моделирования структуры и динамики нуклеосомной фибриллы являются позиции нуклеосом, так как от этого зависят физические характеристики фибриллы, выбор предпочтительной конформации при ее компактизации[7,8]. Популярным методом определения позиций нуклеосом является метод MNase-seq, заключающийся в секвенировании ДНК, оставшейся после гидролиза линкерных участков ферментом микрококковой нуклеазой, однако информация, предоставляемая им, недостаточна для интегративного моделирования, так как не описывает расположение нуклеосом в пространстве относительно друг друга[9]. Решением данной проблемы может быть интеграция данных MNase-seq с данными Micro-C. Micro-C — это модификация метода Hi-C, отличающаяся от стандартных протоколов использованием микрококковой нуклеазы, благодаря чему достигается разрешение, сопоставимое с линейными размерами нуклеосом[10]. Таким образом подход интегративного моделирования структуры и динамики нуклеосомной фибриллы на основании данных Micro-C и MNase-seq является перспективным для изучения организации хроматина на супрануклеосомном уровне.

## 2. ЦЕЛИ И ЗАДАЧИ ИССЛЕДОВАНИЯ

### Цель работы

Целью данной работы является разработка и тестирование методов реконструкции супрануклеосомной структуры и динамики хроматина на основе экспериментальных данных 3D-геномики субнуклеосомного разрешения.

### Задачи

1. Разработать и настроить программные конвейеры для обработки, анализа и интеграции экспериментальных данных Micro-C и MNase-seq
2. С использованием методов огрубленного моделирования ДНК разработать подходы моделирования супрануклеосомной структуры хроматина на основе интегрированных данных
3. С помощью разработанных методов проанализировать супрануклеосомную структуру хроматина в локусе Igf2-H19 для различных клеточных линий *M.musculus* и *H.sapiens*

## 3. ОБЗОР ЛИТЕРАТУРЫ

### 2.1 Уровни компактизации хроматина. Hi-C

ДНК в ядрах эукариотических клеток вместе со специальными белками образует сложный нуклеопротеидный комплекс — хроматин. Хроматин обладает сложной трехмерной структурой, динамически перестраивающейся в соответствии с функциональным состоянием клетки.

В хроматине ДНК может быть компактизована более, чем в $10^5$ раз. Например, геном человека составляет 3 млрд пар оснований, длина одного основания 0,32 нм, т.е длина всей ДНК, находящейся в ядре одной клетки — 1.9 м. При том, что диаметр ядра составляет в среднем около 10 мкм, ДНК оказывается

компактизована в 190000 раз[2]. Расчеты из физики полимеров показывают, что подобная степень компактизации не может быть достигнута по механизму случайного блуждания, без реализации определенных принципов упорядоченного сворачивания — то есть стохастическая укладка хроматина как полимера маловероятна[11]. При этом ДНК в хроматине активно транскрибируется, причем регуляция транскрипции во многом опосредуется структурой хроматина и ее динамическими перестройками[1]. Согласно ряду гипотез, активность таких цис-регуляторных элементов как энхансеры и сайленсеры обусловлена их возможностью физически сближаться с промотором, увеличивая или уменьшая вероятность сборки транскрипционного комплекса для экспрессии соответствующего гена[5,12].

Изучение принципов и закономерностей пространственной организации генома и ее связи с регуляцией транскрипции является одной из основных задач современной молекулярной биологии. Выделяют несколько уровней компактизации хроматина. Высший уровень компактизации хроматина в интерфазном ядре представлен деконденсированными хромосомами, занимающими определенные, неперекрывающиеся области ядра, называемыми хромосомными территориями. Этот уровень был впервые изучен в 1984 году методами флуоресцентной микроскопии[13]. Низший уровень компактизации хроматина — нуклеосомы, соединенные линкерными участками, образующие структуру 10-нанометровой нуклеосомной фибриллы вида “бусины на нити” — был открыт, благодаря электронной микроскопии, в 1974 году, в 1997 году была получена структура нуклеосомы в атомистическом разрешении[14,15].

Промежуточные уровни компактизации хроматина долгое время оставались неисследованными. В экспериментах in vitro была получена и визуализирована с помощью электронной микроскопии структура 30-нанометровой хроматиновой фибриллы, представляющая специфическую укладку 10-нанометровой нуклеосомной фибриллы. Однако позже в серии работ было показано, что 30-нанометровая фибрилла является артефактной структурой, отсутствующей in vivo и не обнаруживающейся в условиях in vitro, близких к

физиологическим, но отличающихся от использованных в экспериментах, в которых она была найдена[16]. Дальнейший прогресс в изучении пространственной структуры генома стал возможен после появления методов 3C — Chromosome Conformation Capture — позволяющих определить, какие локусы с какой частотой физически сближаются в пространстве клеточного ядра. Методы 3C основываются на химической фиксации взаимодействующих в пространстве участков ДНК, фрагментации ДНК, лигировании оставшихся концов фиксированных участков и детекции полученных гибридных молекул с последующим определением их нуклеотидной последовательности, если она не была известна изначально. Первый метод семейства — 3C — позволял за раз получать информацию о взаимодействии только двух локусов с известными нуклеотидными последовательностями, так как детекция гибридной молекулы проводилась с помощью ПЦР. С помощью последующих модификаций протокола — 4C (Chromosome conformation capture on chip), отличающейся использованием микрочипирования после ПЦР, и 5C (Chromosome conformation capture carbon copy), отличающейся использованием мультиплексной ПЦР— можно было определять частоты взаимодействия одного локуса со всеми локусами генома и нескольких выбранных локусов между собой, соответственно[17]. Революционным для трехмерной геномики стал появившийся в 2009 году, основанный на высокопроизводительном секвенировании метод Hi-C, позволяющий определять частоту пространственных взаимодействий всех локусов генома со всеми[18]. Классический протокол Hi-C включает в себя: химическую фиксацию взаимодействующих в пространстве участков ДНК (опосредованно через ассоциированные с ДНК белки), обработку ДНК рестриктазами, биотинилирование концов, лигирование, осаждение полученных гибридных молекул с биотиновой меткой на стрептавидин и последующее секвенирование осажденных гибридных молекул по методу спаренных концов (см. рис.1). Это позволяет получить контактную карту генома — таблицу, в которой каждой паре локусов в соответствие поставлено число (частота) пространственных

контактов между ними, определяющееся числом детектированных гибридных молекул, состоящих из участков этих локусов. Данные Hi-C были использованы для опровержения представления о существовании in vivo регулярной 30-нанометровой фибриллы на супрануклеосомном уровне компактизации хроматина. Также данные Hi-C использовались для аргументации в пользу концепции фрактальной хроматиновой глобулы, согласно которой компактизация хроматина является результатом последовательной сборки глобул все больших порядков, при этом глобулы первого порядка формируются при пространственном контакте относительно близких участков нуклеосомной фибриллы[18]. Кроме этого, благодаря Hi-C, были изучены уровни компактизации хроматина между хромосомным и нуклеосомным. Контактные карты визуализируются тепловыми картами — таблицами, в которых частота пространственных взаимодействий между локусами кодируется интенсивностью выбранного для построения тепловой карты цветом. На тепловых картах Hi-C выявляются определенные паттерны, которые интерпретируются как графическое представление структур, появляющихся на разных уровнях компактизации хроматина. В результате анализа тепловых карт Hi-C были обнаружены такие паттерны, как “шотландская клетка”, или A/B компартменты (этот паттерн детектируется после обработки контактных карт методом главных компонент), и топологически ассоциированные домены (ТАДы) — участки генома, локусы в которых чаще взаимодействуют в пространстве друг с другом, чем с локусами из других участков[19]. Паттерн A/B компартментов считается графическим представлением сегрегации хроматина на активный и неактивный, происходящей на уровне следующим за высшим уровнем компактизации — хромосомным. Паттерн ТАДов интерпретируется как графическое представление уровня компактизации следующего за уровнем компартментов активного и неактивного хроматина, для которого характерно образование глобулярных структур в результате сближения определенных участков фибриллярного хроматина (см.рис.2). В процессе изучения укладки хроматина методом Hi-C было обнаружено существование

ТАДов разных порядков, последовательно вкладывающихся друг в друга[20]. Этот феномен является еще одним доказательством гипотезы фрактальной глобулы.

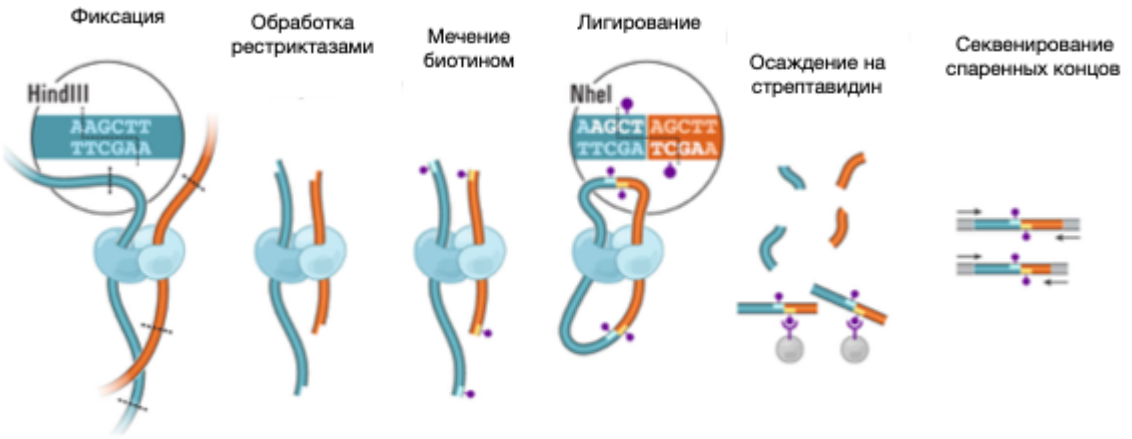


Рис.1. Последовательность действий в классическом протоколе Hi-C. Источник: Lieberman-Aiden et al., 2009

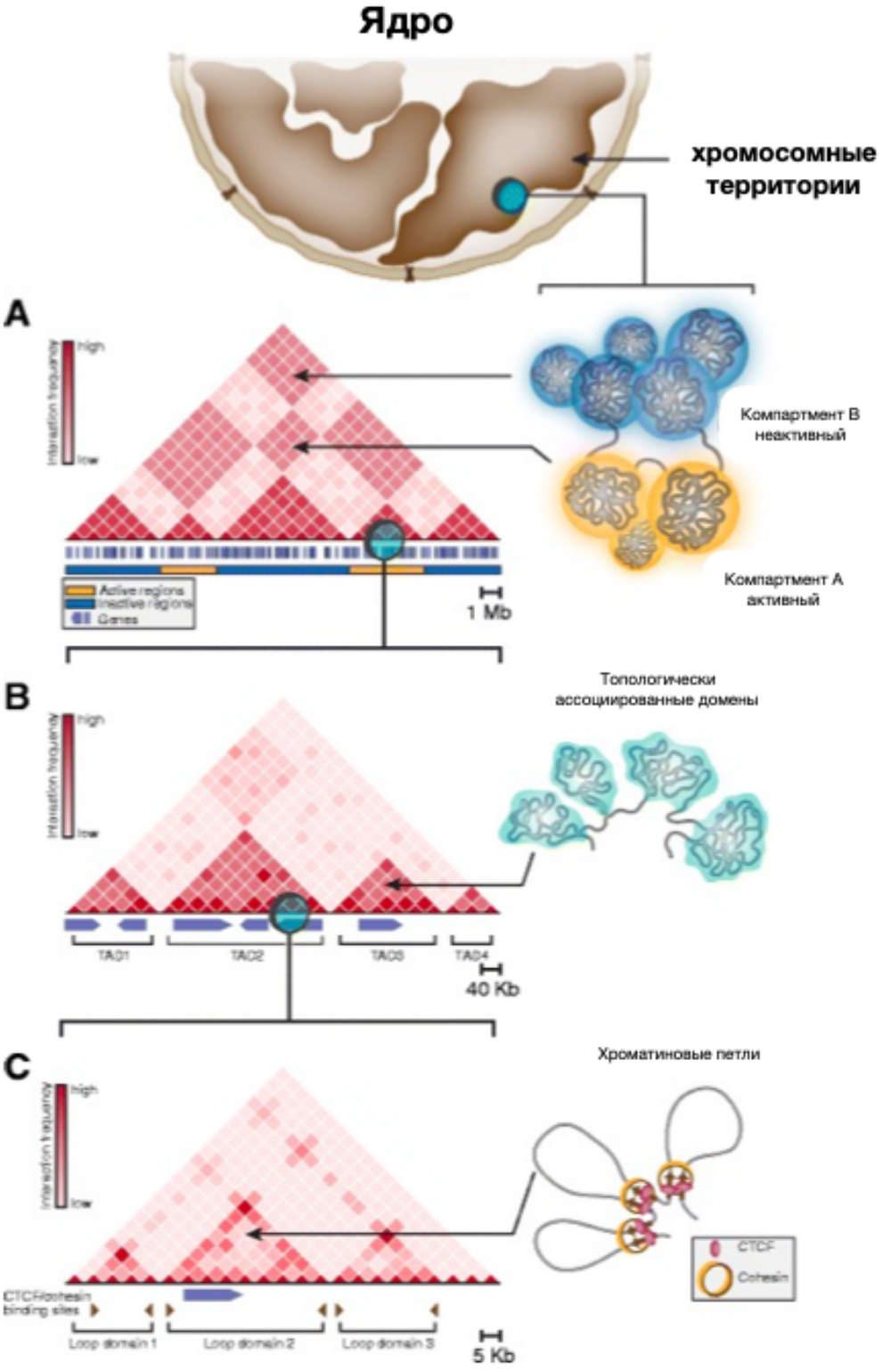


Рис. 2. Уровни компактизации хроматина и паттерны, соответствующие им на тепловых картах Hi-C. Источник: Razin et al., 2017

### 3.2. Механистическая интерпретация паттернов, детектируемых на тепловых картах Hi-C

В рамках одной из наиболее популярных в научном сообществе механистической интерпретации паттернов, детектируемых на тепловых картах Hi-C, структуры, соответствующие A/B компартментам, формируются в процессе микрофазного разделения, а структуры, соответствующие ТАДам — в процессе выпетливания (loop extrusion)[21,22] (см. рис.3). Микрофазное разделение — это процесс образования в блок-сополимерах (полимерах, состоящих из различных структурных сегментов) глобулярных структур в результате сближения сегментов одного типа[23]. Эффект микрофазного разделения наблюдается в экспериментах с нуклеосомными фибриллами in vitro, так же как и аналогичный для модельных нуклеосомных олигомеров эффект разделения фаз по типу жидкость-жидкость (liquid-liquid phase separation). Эти эффекты объясняются прежде всего разницей в зарядах гистоновых хвостов нуклеосом (определяемой во многом статусом ацетилирования гистонового октамера в позиции H4K16): часть нуклеосом электростатически притягивается друг к другу, благодаря взаимодействию отрицательно заряженного кислотного лоскутка одной нуклеосомы и положительно заряженного гистонового хвоста другой, а часть нуклеосом не притягивается из-за того, что гистоновый хвост ацетилирован[24]. В пользу этой модели говорит то, что ацетилирование гистонового хвоста, предотвращающее взаимодействие нуклеосом, ассоциировано с активными компартментами хроматина[25]. Выявляются также и другие параметры, влияющие на процесс микрофазного разделения, например, кратность длины линкерных участков[26]. Процесс микрофазного разделения в хроматине, согласно современным представлениям, приводит к сегрегации активных и неактивных участков хроматина в обособленные компартменты. Выпетливание (loop extrusion) — это гипотетический процесс, происходящий в хроматине и заключающийся в формировании на нуклеосомной фибрилле петли, сближающей промоторы и цис-регуляторные элементы[27]. Такая петля

формируется в результате активной загрузки на нуклеосомную фибриллу кольцевой молекулы когезина и ее движения по формирующейся петле до барьера, создаваемого белком CTCF, связывающегося с ДНК в специфических сайтах, фланкирующих участок, из которого формируется петля[28]. В пользу модели “loop extrusion” говорят результаты множества исследований[29].

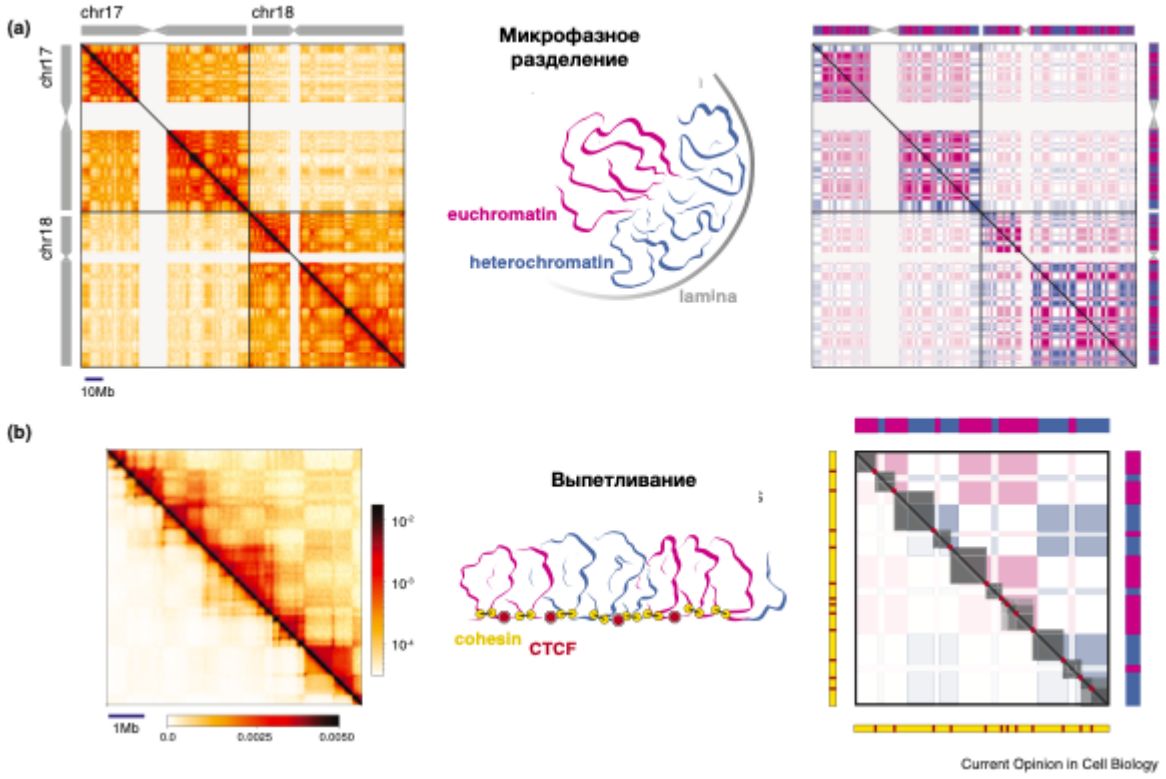


Рис.3 Процессы выпетливания и микрофазного разделения в схематическом представлении и соответствующие им паттерны на тепловых картах Hi-C. Источник: Mirny et al., 2019

### 3.3. Разрешающая способность метода Hi-C. Micro-C.

Под разрешением (разрешающей способностью) метода Hi-C понимают минимальные размеры локусов, между которыми определяется частота пространственных взаимодействий. Разрешение соответствует покрытию прочтениями, приходящимися на фрагмент ДНК. Таким образом, разрешение Hi-C определяется такими факторами, как глубина секвенирования и тип рестриктазы, выбранной для фрагментации ДНК после химической фиксации

пространственно взаимодействующих участков[19]. Влияние выбора рестриктаз на разрешение Hi-C обусловлено размерами сайтов рестрикции и их распределением в геноме. Чем больше сайт рестрикции, тем статистически реже он встречается в геноме, тем больше будут фрагменты ДНК после обработки рестриктазами и тем меньше будет разрешение. Неравномерное распределение сайтов рестрикции приводит к неравномерному распределению покрытия относительно фрагмента и также отрицательно сказывается на разрешении. В первом протоколе Hi-C использовались рестриктазы с сайтами рестрикции размерами 6 п.о. (HindIII, NcoI), что позволяло достигать разрешения порядка 1 мегабаз, в последующих модификациях протокола начали использоваться рестриктазы с сайтами рестрикции в 4 п.о. (DpnII, MboI), что позволило улучшить разрешение до нескольких килобаз[19]. Однако, так как длина нуклеосомной ДНК составляет обычно 146 п.о., увеличение разрешения даже до нескольких килобаз не позволяет изучать супрануклеосомный уровень компактизации хроматина. Изучение непосредственно супрануклеосомного уровня компактизации хроматина стало возможным после появления Micro-C — модификации Hi-C, которая за счет использования микрококковой нуклеазы вместо рестриктаз позволяет достигать субнуклеосомного разрешения — разрешения порядка 100-200 п.о, сопоставимого с линейными размерами нуклеосом[10]. Микрококковая нуклеаза — фермент, обладающий как эндо-, так и экзонуклеазной активностью и за счет этого гидролизующий большую часть линкерной ДНК. Помимо Micro-C существуют и другие модификации Hi-C, не основанные на использовании рестриктаз, например, DNase-HiC, в котором используется днказа или ChIA-PET, в котором перед проведением Hi-C проводится иммунопреципитация хроматина на белок, ассоциированный с ДНК в определенном локусе[30,31]. Эти методы позволяют значительно повысить разрешение данных, однако разрешение, достигаемое с помощью Micro-C, выше, чем в методе DNAse-HiC, а ChIA-PET возможно использовать только для определенных локусов, а не для всего генома.

### 3.4. Подходы к моделированию структуры и динамики нуклеосомной фибриллы

Так как нуклеосомная фибрилла представляет собой полимер, описываемый набором физических параметров,  таких как эластичность линкерных участков, электростатический потенциал нуклеосом и линкерных участков, объем нуклеосом и т.д, её компактизацию — и, соответственно супрануклеосомную структуру хроматина — можно моделировать методами молекулярной динамики, находя оптимальные значения суммарной энергии фибриллы. Для молекулярного моделирования фибриллы было разработано множество подходов[6,32–35]. Большая часть из них представляет собой методы огрубленного моделирования, в которых нуклеосомы представляются в виде простых тел — цилиндров или шаров — а линкерные участки в виде пружинных или жестких связок[36]. Сильная сторона подобных подходов заключается в простоте вычислений, однако игнорирование ряда параметров (например, вариативности длины линкера) может привести к получению недостоверной модели, так как игнорируемые параметры могут оказаться важными для процесса компактизации фибриллы. Более гибким подходом, считающимся некоторыми авторами более эффективным для моделирования нуклеосомной фибриллы, является мезоскопический подход, в котором ДНК представляется не в виде цепочки пружин или жестких связок, соответствующим линкерным участкам, а в виде дискретных единиц соответствующих нуклеотидным парам и обладающим несколькими степенями свободы[37]. Популярным методом моделирования, использующим мезоскопический подход, является метод динуклеотидного приближения ДНК. Этот метод заключается в представлении ДНК в виде последовательности динуклеотидных “шагов”, траектория которой определяется шестью параметрами: Tilt (наклон), Roll (качение), Twist (поворот), Shift (сдвиг), Slide (скольжение), Rise (подъем)[38]. Использование этого метода удобно еще и возможностью применения вместе с ним подходов оптимизации энергии, основанных на принципах Монте-Карло: шагом симуляции является случайный

выбор пары оснований и изменение значений характеризующих ее параметров[39]. После такого случайного изменения проводится измерение разницы в энергии между новой и предыдущей конформацией нуклеосомной фибриллы и переход к новой конформации в случае, если она оказывается ближе к энергетическому оптимуму. Для любого метода моделирования достоверность результатов зависит от того, как будут заданы входные параметры, поэтому для эффективного применения моделирования как способа реконструкции структуры хроматина необходимо основывать его на экспериментальных данных — то есть использовать интегративный подход. Примером интегративного подхода к изучению супрануклеосомной структуры хроматина является работа Ohno et al., 2019[40], в рамках которой был разработан метод Hi-Co, представляющий собой моделирование структуры нуклеосомной фибриллы хроматина дрожжей *C.cerevisiae* на основании данных, полученных с помощью модифицированного протокола Micro-C, позволяющего локализовать нуклеосомы и получить информацию об их ориентации на нуклеосомной фибрилле. Недостатком метода Hi-Co является относительная сложность экспериментальной части и отсутствие опыта использования этого метода другими группами исследователей. В результате нельзя сделать однозначных выводов относительно воспроизводимости результатов, полученных с помощью Hi-Co, и соответственно, перспективности этого метода для изучения супрануклеосомного уровня компактизации хроматина.

### 3.5. Информация о позиционировании нуклеосом. MNase-seq

Важным параметром, который необходимо задать для реалистичного моделирования структуры и динамики нуклеосомной фибриллы являются позиции нуклеосом, так как физические свойства нуклеосомной фибриллы могут значительно изменяться в зависимости от плотности расположения нуклеосом, которые при этом могут определенным образом взаимодействовать — например, электростатически, что может приводить, как было описано выше,

к глобальным конформационным перестройкам по механизму микрофазного разделения. Для определения позиций нуклеосом был разработан ряд экспериментальных техник: ATAC-seq (transposase), ChIP-seq (H3 histone), NOME-seq, MRE-seq, FAIRE-seq и др.[41]. Одним из наиболее популярных из них является  MNase-seq[9]. Преимущество MNase-seq заключается в том, что в отличие от большинства других методов он основан не на детекции ассоциированного с нуклеосомами белка, а на непосредственном секвенировании нуклеосомной ДНК, остающейся в реакционной смеси после ее обработки микрококковой нуклеазой, которая обладает способностью к гидролизу линкерных участков. После биоинформатической обработки данные MNase-seq  могут быть представлять собой дискретный набор координат диад (центров) нуклеосом (positioning), а могут быть непрерывным сигналом нуклеосомной занятости (occupancy), соответствующим над каждой парой оснований части клеточной популяции, у которой эта пара оснований взаимодействует с нуклеосомой[42]. Так как для моделирования нуклеосомной фибриллы требуется задать точные положения нуклеосом, данные нуклеосомной занятости перед их использованием для задания параметров моделирования необходимо трансформировать в данные позиционирования нуклеосомных диад. Задание позиций нуклеосом является важным условием для реалистичного моделирования структуры и динамики нуклеосомной фибриллы, однако недостаточным, так не менее важным, чем позиции нуклеосом, параметром является их взаимное расположение, информацию о котором метод MNase-seq не предоставляет. Решением данной проблемы может быть интеграция данных MNase-seq с данными Micro-C. Интеграция таких данных о пространственных взаимодействиях в геноме с данными о позициях нуклеосом может обеспечить информацию о пространственных контактах между нуклеосомами, необходимую для реалистичного моделирования структуры и динамики нуклеосомной фибриллы, компактизующейся в супрануклеосомную структуру.

## 4. МЕТОДЫ И МАТЕРИАЛЫ ИССЛЕДОВАНИЯ

Молекулярное моделирование структуры нуклеосомной фибриллы, разработка и настройка программного обеспечения проводилась на кластере Newton. Разработка программного обеспечения проводилась на таких языках программирования как Python 3, R, bash в интегрированной среде разработки JupyterLab.

### 4.1. Обработка данных Micro-C

Для обработки данных Micro-C на вычислительном кластере, на котором проводились необходимые вычисления и разработка программного обеспечения, был настроен программный конвейер distiller (https://github.com/dekkerlab/distiller-nf), установлены необходимые для его работы программные библиотеки: pairtools (https://pairtools.readthedocs.io/en/latest/), cooler (https://cooler.readthedocs.io/en/latest/index.html). Последовательная обработка данных Micro-C с помощью программного конвейера  distiller организована следующим образом. На вход подаются данные секвенирования в формате FASTAQ. Первым этапом их обработки является выравнивание прочтений на референсный геном с помощью биоинформатического инструмента BWA[43], реализующего алгоритм выравнивания Смита-Ватермана. Результат выравнивания сохраняется в бинарном формате BAM. На следующем этапе происходит фильтрация ПЦР-дупликатов и неоднозначных выравниваний. После этого с помощью программной библиотеки pairtools среди выровненных

последовательностей находятся последовательности, соответствующие концам гибридных молекул, образовавшихся в процессе лигирования фиксированных участков ДНК, пространственно взаимодействующих в хроматине. Список детектированных таким образом пространственных взаимодействий в хроматине сохраняется в формате PAIRS. На заключительном этапе обработки данных Micro-C с помощью программного конвейера distiller из списка пар формируется контактная карта с заданным исследователем разрешением, если оно может быть достигнуто при данной глубине секвенирования. Контактные карты записываются в формате COOL и представляют собой список контактирующих локусов с указанием значения частоты их пространственного взаимодействия (числа детекций данного контакта).

**4.2. Молекулярное моделирование**

Для моделирования структуры нуклеосомной фибриллы использовалась разрабатываемая в группе интегративной биологии программная библиотека PyNaMod. Данная библиотека сочетает подход огрубленного моделирования нуклеосомной фибриллы, в рамках которого нуклеосомы представляются шарами, и мезоскопический подход, в рамках которого используется динуклеотидное приближение ДНК: линкерные участки представляются цепочками прямоугольников, каждая из которых соответствует паре оснований, их взаиморасположение определяют шесть параметров: Tilt (наклон), Roll (качение), Twist (поворот), Shift (сдвиг), Slide (скольжение), Rise (подъем) -- см.рис.4. Процесс моделирования структуры фибриллы основан на принципах Монте-Карло и представляет собой итеративное изменение параметров случайных динуклеотидных пар, сопровождающееся проверкой изменения энергии фибриллы после каждого изменения параметров и закреплением изменения параметров в случае превышения измеренной разницей энергий порогового значения. Данный процесс продолжается до тех пор, пока не достигается глобальный энергетический оптимум моделируемой системы, критерием чего является выхождение функции полной энергии фибриллы на плато.

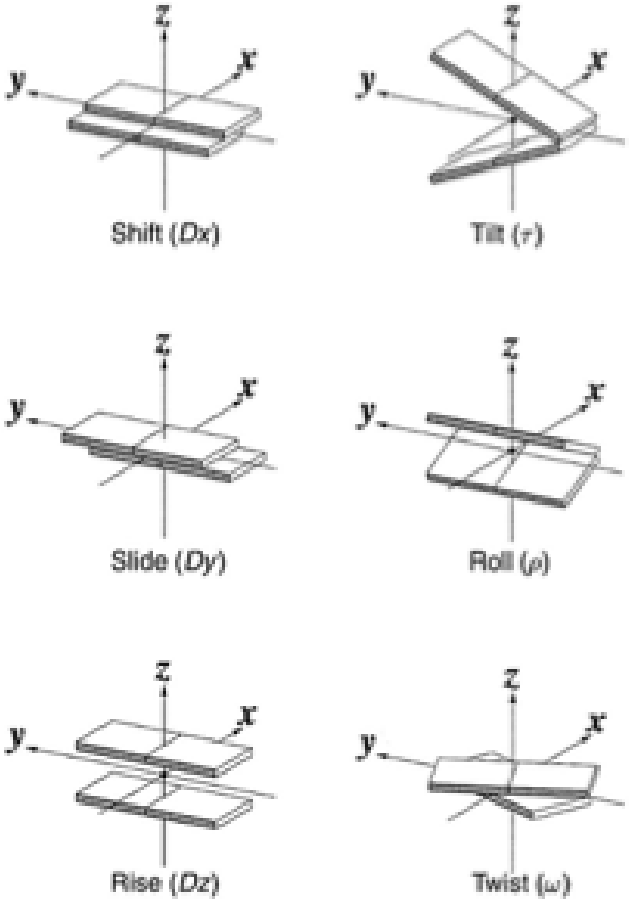


Рис.4 Параметры пар оснований в динуклеотидном приближении ДНК. Источник: Xiang-Jun Lu et al., 2003

### 4.3. Использованные репозитории и базы данных. Данные, выбранные для тестирования разработанного программного конвейера

Для тестирования разработанного в ходе работы программного конвейера использовались данные Micro-C, размещенные в открытом репозитории 4DNucleome, и данные MNase-seq, полученные с помощью репозитория NucPosDB, представляющего собой агрегатор ссылок на базу данных GEO (см. Приложение, Таблица 1). Также в работе использовалась база данных Ensembl для получения данных нуклеотидных последовательностей, необходимых для моделирования нуклеосомной фибриллы.

В рамках тестирования разработанного в ходе работы программного конвейера была проведена интеграция данных Micro-C и MNase-seq для клеточных линий

*H.sapiens:* H1-hESC, HeLa-S3 и *M.musculus:* JM8.N4 (mESC). В виду ограничений вычислительной мощности разработанного программного конвейера проводился анализ отдельного локуса, а не полного генома. Для анализа был выбран локус Igf2-H19 (с геномными координатами для *M.musculus* — 7:142,460,685-142,720,685 и для *H.sapiens* — 11:1,911,877-2,181,877) так как он является одним из наименьших локусов, для которых в ходе ряда исследований было доказано участие процесса петлеобразования в регуляции транскрипции[44].

## 5. РЕЗУЛЬТАТЫ РАБОТЫ И ОБСУЖДЕНИЕ

В ходе работы был разработан и настроен программный конвейер для интеграции данных Micro-C и MNase-seq (см.рис.5). Работу конвейера можно условно разделить на три этапа. На первом этапе данные Micro-C обрабатываются с помощью программного конвейера distiller, проходя путь от данных секвенирования в формате FASTQ до контактных карт с разрешением 200 п.о. в формате COOL и выгружаясь в виде таблицы в текстовом формате BEDPE. На втором этапе данные MNase-seq преобразуются из данных нуклеосомной занятости (occupancy) в бинарном формате WIG в данные позиций диад нуклеосом в текстовом формате BED. Если данные MNase-seq изначально подаются на вход программного конвейера в виде данных о позициях диад нуклеосом в текстовом формате BED, то на втором этапе они остаются без изменений. Разработанный в ходе работы алгоритм преобразования данных нуклеосомной занятости (occupancy) в данные позиций диад нуклеосом следующий: устанавливается пороговое значение сигнала нуклеосомной занятости, отбираются участки, в которых сигнал превосходит данное значение, фиксируются пики сигнала и вычисляются середины оснований пиков, координаты которых заносятся в список координат диад нуклеосом. Полученный список координат диад

нуклеосом фильтруется таким образом, чтобы расстояние между соседними пиками не было меньше, чем 146 п.о., что обусловливается необходимостью исключить возможность пересечения нескольких нуклеосом моделируемой нуклеосомной фибриллы в одной точке пространства. Третьим этапом работы программного конвейера является интеграция данных контактных карт Micro-C с разрешением 200 п.о. с данными позиций диад нуклеосом. На данном этапе в соответствии каждой диаде ставится локус контактной карты, на который данная диада картируется. С помощью этого соответствия создается нуклеосомная контактная карта, представляющая собой список контактов между диадами нуклеосом, в котором значение частоты пространственного взаимодействия между парой нуклеосом равно значению частоты пространственного взаимодействия между локусами контактной карты Micro-C, на которые картируются диады данных нуклеосом. Для подготовки к моделировании структуры нуклеосомной фибриллы на основе полученных нуклеосомных контактных карт в разработанный программный конвейер включена возможность перевода значений частот пространственных взаимодействий между нуклеосомами в значения дистанций между нуклеосомами по формуле $D = 1/x^a$, (где D -- дистанция в условных единицах, x -- частота взаимодействий) применимость которой для ряда значений была показана в ряде работ[45]. В ходе тестирования разработанного программного конвейера значение $a$ бралось равным единице согласно работе Lieberman-Aiden, 2009[18].

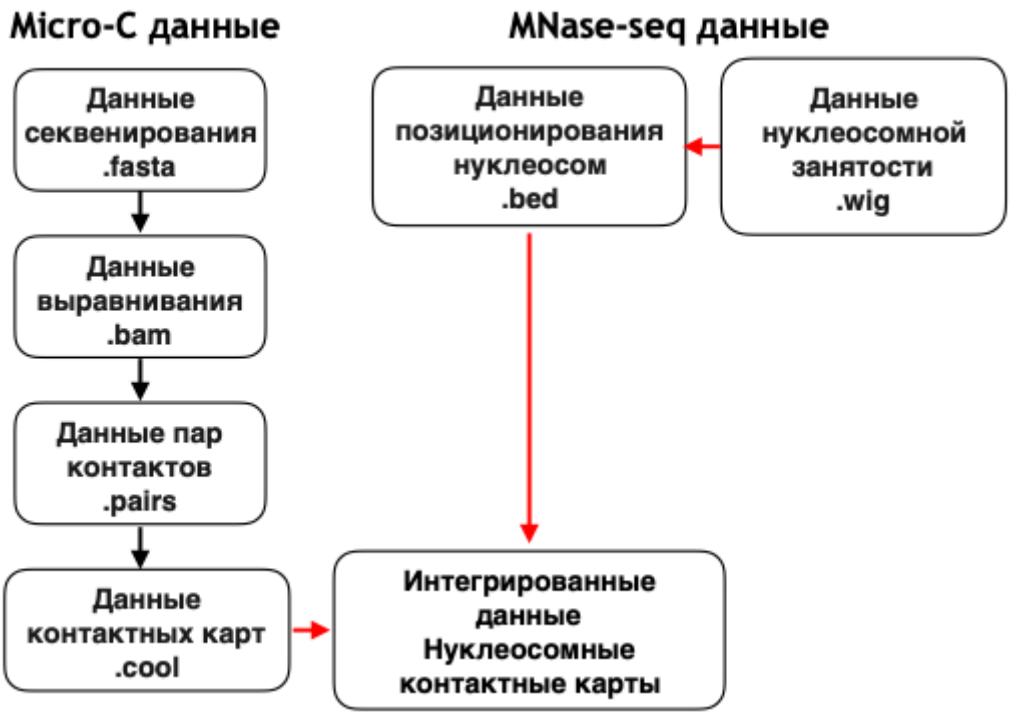


Рис.5. Блок-схема разработанного программного конвейера

В ходе работы были построены контактные нуклеосомные карты для локуса Igf2-H19 для клеточных линий *H.sapiens:* H1-hESC, HeLa-S3 и *M.musculus:* JM8.N4 (mESC), см.рис.6-8. Ввиду отсутствия специального программного обеспечения для данных подобного рода детекция паттернов проводилась только на уровне визуального анализа, а выявление сходства проводилось с помощью IMGonline — неспециализированного приложения для анализа изображений, находящегося в открытом доступе (https://www.imgonline.com.ua). Сопоставительный анализ выявил сходство нуклеосомных контактных карт, полученных для разных клеточных линий одного биологического вида — для линий H1-hESC, HeLa-S3 *H.sapiens* — коэффициент сходства, рассчитанный с помощью IMGonline и соответствующий доле совпадающих пикселей, составил 88,27%. При этом сопоставительный анализ не выявил сходства между нуклеосомными контактными картами для разных биологических видов — H1-hESC *H.sapiens* и JM8.N4 *M.musculus* — в данном случае коэффициент сходства, рассчитанный с помощью IMGonline, составил 21,68%. Визуальный анализ всех полученных нуклеосомных карт локуса Igf2-H19 выявил на них

ТАД-подобные паттерны, которые, согласно современным представлениям об организации хроматина у млекопитающих, ассоциированы с процессом выпетливания (loop extrusion).[21,22]

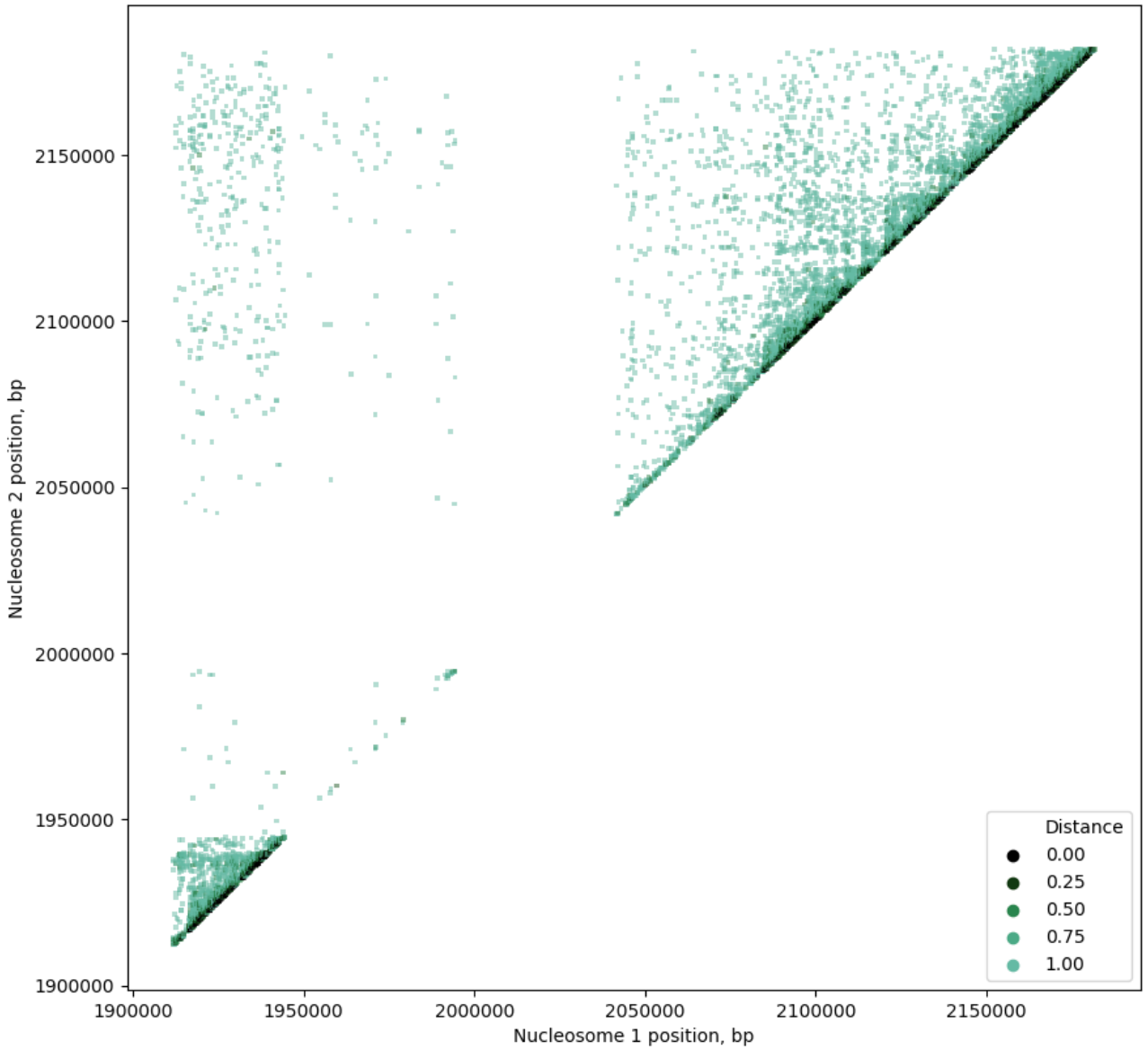


Рис.6. Нуклеосомная контактная карта локуса Igf2-H19 *H.sapiens,* клеточная линия H1-hESC

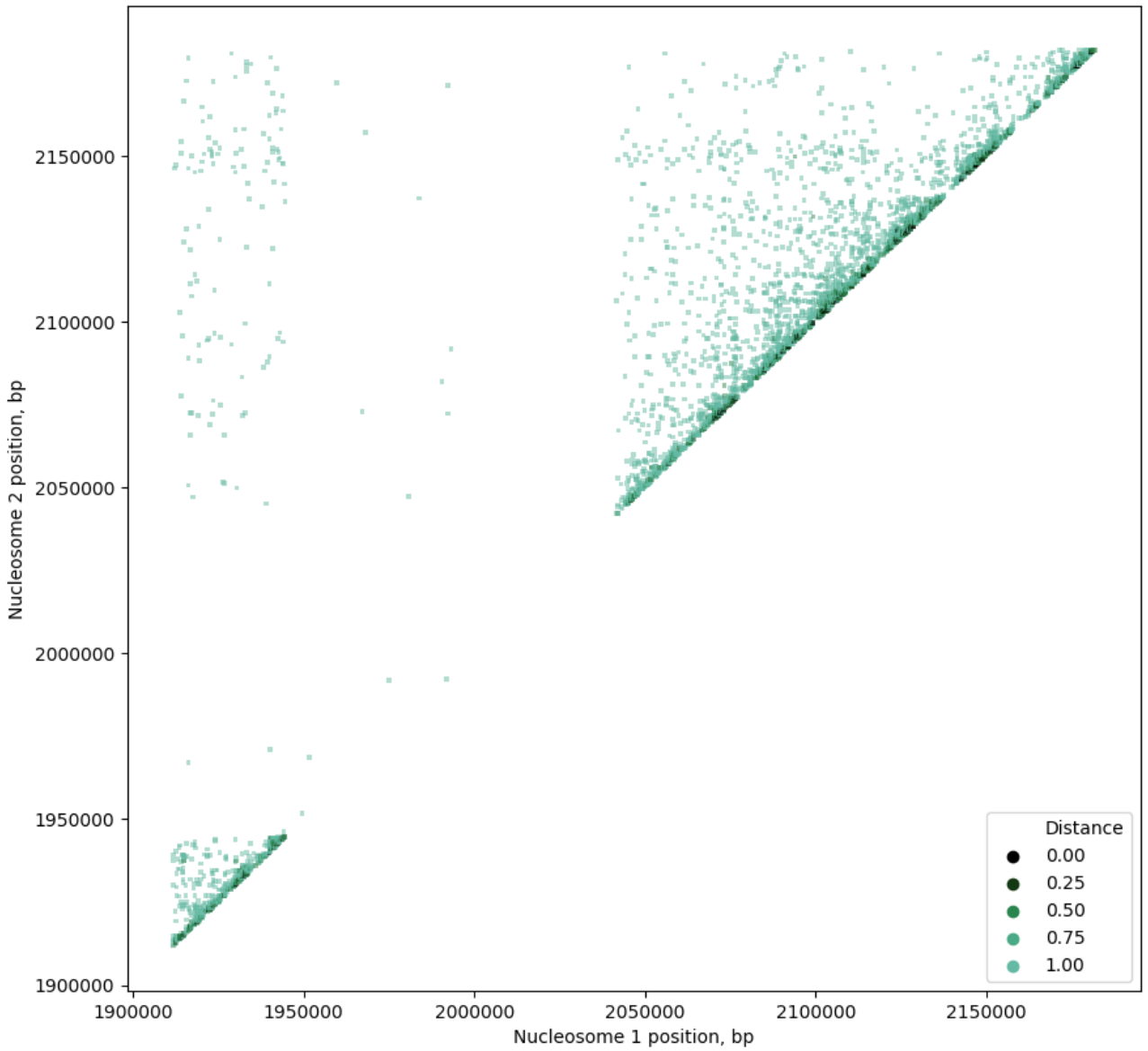


Рис.7. Нуклеосомная контактная карта локуса Igf2-H19 *H.sapiens,* клеточная линия HeLa-S3

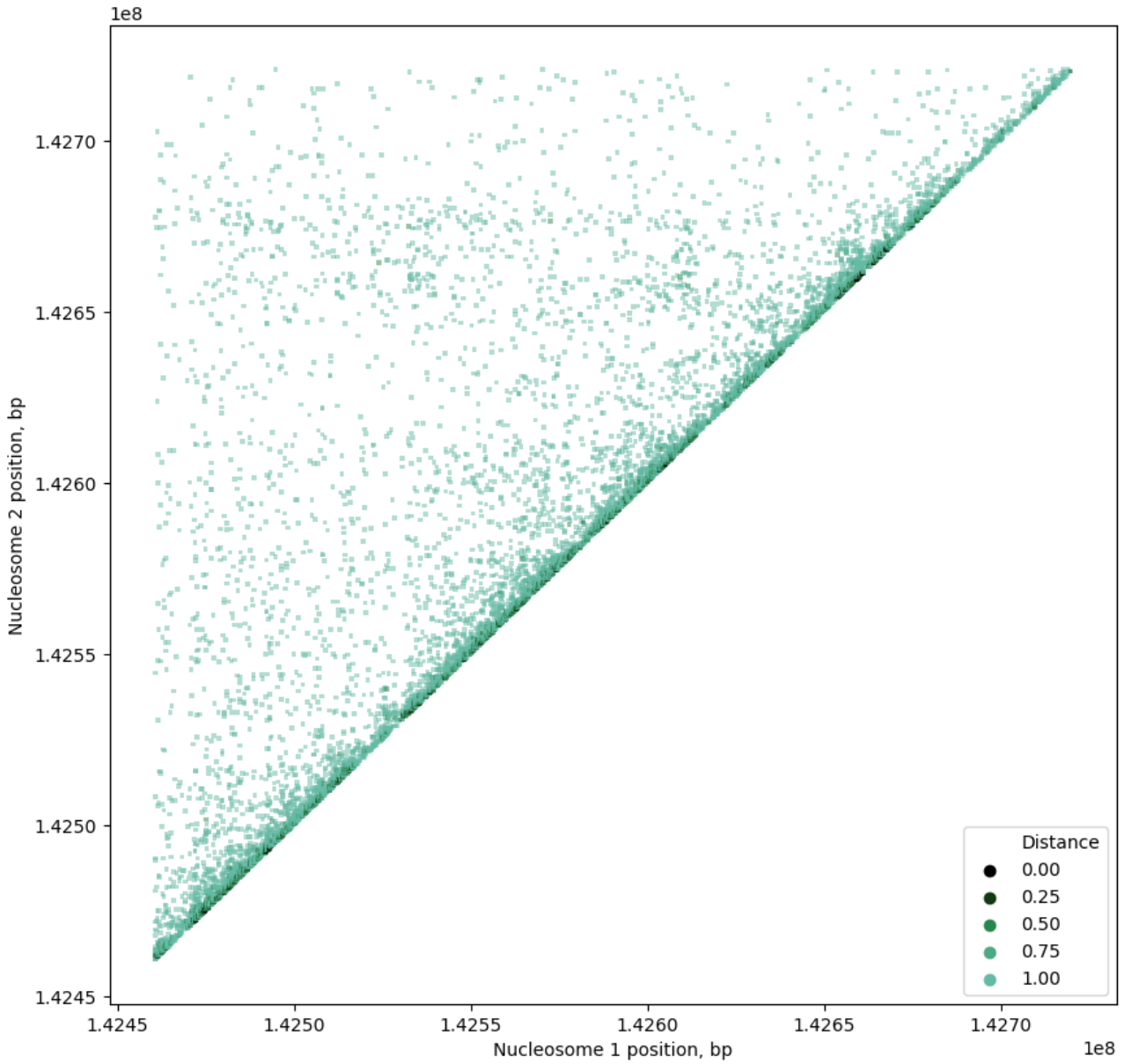


Рис.8. Нуклеосомная контактная карта локуса Igf2-H19 *M.musculus*, клеточная линия JM8.N4

В ходе работы с помощью программной библиотеки PyNaMod была проведена реконструкция супрануклеосомной структуры хроматина путем молекулярного моделирования компактизации нуклеосомной фибриллы на основании полученных с помощью разработанного программного конвейера интегрированных данных MNase-seq и Micro-C (см.рис.9). Ввиду ограниченности вычислительной мощности программной библиотеки PyNaMod для моделирования был выбран участок локуса IgF2-H19 *M.musculus* размерами 10 kb. Реконструированный участок представляет собой нерегулярно

организованную фибриллу, периодически образующую за счет дополнительной спирализации глобулярные структуры. Полученные данные согласуются с современными представлениями о нерегулярности организации хроматина на супрануклеосомном уровне.

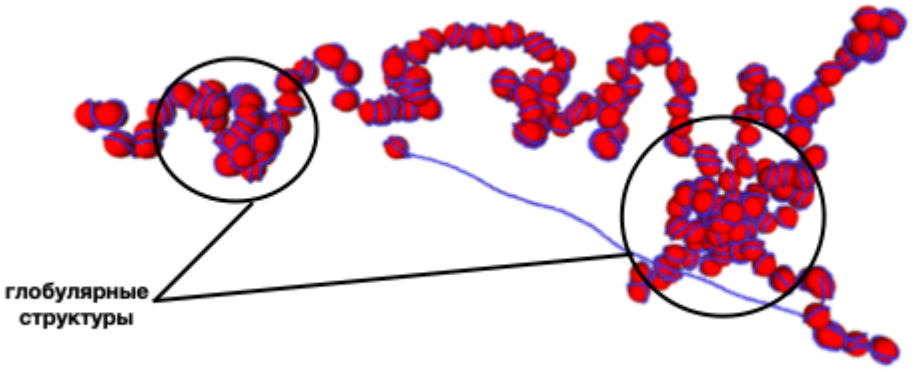


Рис.9. Реконструкция супрануклеосомной структуры хроматина в участке локуса Igf2-H19 *M.musculus*, клеточная линия JM8.N4

## 6. ЗАКЛЮЧЕНИЕ

В ходе данной работы был разработан и настроен программный конвейер для интеграции данных Micro-C и MNase-seq, получению на их основе нуклеосомных контактных карт и моделированию на их основе структуры нуклеосомной фибриллы с помощью программной библиотеки PyNaMod. В рамках тестирования разработанного программного конвейера были получены нуклеосомные контактные карты локуса Igf2-H19 для клеточных линий *H.sapiens:* H1-hESC, HeLa-S3 и *M.musculus:* JM8.N4 (mESC). Сопоставительный анализ полученных контактных нуклеосомных карт показал сходство для линий *H.sapiens* H1-hESC и HeLa-S3 и различие для линий *H.sapiens H1-hESC* и JM8.N4 *M.musculus,* что может указывать на значимость фактора видовой принадлежности для организации супрануклеосомного уровня компактизации хроматина в данном локусе. Визуальный анализ всех полученных нуклеосомных карт выявил на них ТАД-подобные паттерны, В ходе данной

работы также было проведено моделирование с помощью программной библиотеки PyNaMod структуры нуклеосомной фибриллы для участка локуса IgF2-H19 *M.musculus.* Моделирование показало, что на супрануклеосомном уровне хроматин представляет нерегулярно организованную фибриллу, локально образующую глобулярные структуры. Эти результаты согласуются с современными представлениями об организации хроматина на супрануклеосомном уровне.

Таким образом нами был разработан и протестирован на экспериментальных данных программный конвейер для интеграции данных Micro-C и MNase-seq и последующего молекулярного моделирования структуры нуклеосомной фибриллы. Полученные в ходе тестирования разработанного программного конвейера данные согласуются с современными представлениями о нерегулярности организации хроматина на супрануклеосомном уровне.

## 7. **ВЫВОДЫ**

1. В ходе работы были разработаны и настроены программные конвейеры, обрабатывающие данные Micro-C и интегрирующие их с данными MNase-seq для получения нуклеосомных контактных карт
2. Были созданы протоколы моделирования, позволяющие на основе подходов Монте-Карло и моделирования ДНК в динуклеотидном приближении моделировать на основании нуклеосомных карт супрануклеосомную структуру хроматина
3. На основании данных Micro-C и MNase-seq для клеточных линий *H.sapiens:* H1-hESC, HeLa-S3 и *M.musculus:* JM8.N4 (mESC) были вычислены нуклеосомные контактные карты локуса Igf2-H19.
4. В ходе визуального анализа полученных нуклеосомных карт выявлено наличие ТАД-подобных паттернов
5. Сравнительный анализ полученных нуклеосомных контактных карт локуса Igf2-H19 для клеточных линий H1-hESC, HeLa-S показал их сходство.

6. Сравнительный анализ полученных нуклеосомных контактных карт локуса Igf2-H19 *H.sapiens* и *M.musculus* показал отсутствие значимого сходства между ними.

7. Реконструкция участка супрануклеосомной структуры хроматина в локусе Igf2-H19 *M.musculus* представляет собой нерегулярно организованную, локально образующую глобулярные структуры, фибриллу

## 8. СПИСОК ЛИТЕРАТУРЫ


1. Cavalli G., Misteli T. Functional implications of genome topology // Nat. Struct. Mol. Biol. 2013. Vol. 20, № 3. P. 290–299.
2. Razin S.V., Ulianov S.V. Gene functioning and storage within a folded genome // Cell. Mol. Biol. Lett. 2017. Vol. 22. P. 18.
3. Kornberg R.D. Chromatin structure: a repeating unit of histones and DNA // Science. American Association for the Advancement of Science, 1974. Vol. 184, № 4139. P. 868–871.
4. Fraser J. et al. An Overview of Genome Organization and How We Got There: from FISH to Hi-C // Microbiol. Mol. Biol. Rev. MMBR. American Society for Microbiology, 2015. Vol. 79, № 3. P. 347–372.
5. Holwerda S., de Laat W. Chromatin loops, gene positioning, and gene expression // Front. Genet. 2012. Vol. 3.
6. Ozer G., Luque A., Schlick T. The Chromatin Fiber: Multiscale Problems and Approaches // Curr. Opin. Struct. Biol. 2015. Vol. 31. P. 124–139.
7. Nikitina T. et al. DNA topology in chromatin is defined by nucleosome spacing // Sci. Adv.

2017. Vol. 3, № 10. P. e1700957.
8. Norouzi D. et al. Topological diversity of chromatin fibers: Interplay between nucleosome repeat length, DNA linking number and the level of transcription // AIMS Biophys. 2015. Vol. 2, № 4. P. 613–629.
9. Klein D.C., Hainer S.J. Genomic methods in profiling DNA accessibility and factor localization // Chromosome Res. 2020. Vol. 28, № 1. P. 69–85.
10. Hsieh T.-H.S. et al. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C // Cell. 2015. Vol. 162, № 1. P. 108–119.
11. Mirny L.A. The fractal globule as a model of chromatin architecture in the cell // Chromosome Res. 2011. Vol. 19, № 1. P. 37–51.
12. Kadauke S., Blobel G.A. Chromatin loops in gene regulation // Biochim. Biophys. Acta. 2009. Vol. 1789, № 1. P. 17–25.
13. Cremer T., Cremer M. Chromosome Territories // Cold Spring Harb. Perspect. Biol. 2010. Vol. 2, № 3.
14. Kornberg R.D., Thomas J.O. Chromatin structure; oligomers of the histones // Science. 1974. Vol. 184, № 4139. P. 865–868.
15. Luger K. et al. Crystal structure of the nucleosome core particle at 2.8 Å resolution // Nature. Macmillan Magazines Ltd., 1997. Vol. 389. P. 251–251.
16. Razin S.V., Gavrilov A.A. Chromatin without the 30-nm fiber // Epigenetics. 2014. Vol. 9, № 5. P. 653–657.
17. Han J., Zhang Z., Wang K. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering // Mol. Cytogenet. 2018. Vol. 11, № 1. P. 21.
18. Lieberman-Aiden E. et al. Comprehensive mapping of long range interactions reveals folding principles of the human genome // Science. 2009. Vol. 326, № 5950. P. 289–293.
19. Pal K., Forcato M., Ferrari F. Hi-C analysis: from data generation to integration // Biophys. Rev. 2019. Vol. 11, № 1. P. 67–78.
20. Phillips-Cremins J.E. et al. Architectural protein subclasses shape 3-D organization of genomes during lineage commitment // Cell. 2013. Vol. 153, № 6. P. 1281–1295.
21. Mirny L.A., Imakaev M., Abdennur N. Two major mechanisms of chromosome organization // Curr. Opin. Cell Biol. 2019. Vol. 58. P. 142–152.
22. Nuebler J. et al. Chromatin organization by an interplay of loop extrusion and compartmental segregation // Proc. Natl. Acad. Sci. U. S. A. 2018. Vol. 115, № 29. P. E6697–E6706.
23. Razin S.V., Ulianov S.V. Divide and Rule: Phase Separation in Eukaryotic Genome Functioning // Cells. 2020. Vol. 9, № 11.
24. Allahverdi A. et al. The effects of histone H4 tail acetylations on cation-induced chromatin folding and self-association // Nucleic Acids Res. 2011. Vol. 39, № 5. P. 1680–1691.
25. Agalioti T., Chen G., Thanos D. Deciphering the transcriptional histone acetylation code for a human gene // Cell. 2002. Vol. 111, № 3. P. 381–392.
26. Gibson B.A. et al. Organization of Chromatin by Intrinsic and Regulated Phase Separation // Cell. 2019. Vol. 179, № 2. P. 470-484.e21.
27. Fudenberg G. et al. Formation of Chromosomal Domains by Loop Extrusion // Cell Rep. 2016. Vol. 15, № 9. P. 2038–2049.
28. Davidson I.F. et al. DNA loop extrusion by human cohesin // Science. 2019. Vol. 366, № 6471. P. 1338–1345.
29. Fudenberg G. et al. Emerging Evidence of Chromosome Folding by Loop Extrusion // Cold Spring Harb. Symp. Quant. Biol. 2017. Vol. 82. P. 45–55.
30. Ramani V. et al. Mapping 3D genome architecture through in situ DNase Hi-C // Nat. Protoc. 2016. Vol. 11, № 11. P. 2104–2121.
31. Fullwood M.J. et al. An Oestrogen Receptor α-bound Human Chromatin Interactome // Nature. 2009. Vol. 462, № 7269. P. 58–64.
32. Varoquaux N. et al. A statistical approach for inferring the 3D structure of the genome // Bioinforma. Oxf. Engl. 2014. Vol. 30, № 12. P. i26-33.

33. Adhikari B., Trieu T., Cheng J. Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing // BMC Genomics. 2016. Vol. 17, № 1. P. 886.
34. Rousseau M. et al. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling // BMC Bioinformatics. 2011. Vol. 12. P. 414.
35. Schlick T. Monte Carlo, harmonic approximation, and coarse-graining approaches for enhanced sampling of biomolecular structure // F1000 Biol. Rep. 2009. Vol. 1.
36. Korolev N., Nordenskiöld L., Lyubartsev A.P. Multiscale coarse-grained modelling of chromatin components: DNA and the nucleosome // Adv. Colloid Interface Sci. 2016. Vol. 232. P. 36–48.
37. Olson W.K. et al. Influence of fluctuations on DNA curvature. A comparison of flexible and static wedge models of intrinsically bent DNA // J. Mol. Biol. 1993. Vol. 232, № 2. P. 530–554.
38. Dickerson R.E. Definitions and nomenclature of nucleic acid structure components // Nucleic Acids Res. 1989. Vol. 17, № 5. P. 1797–1803.
39. Norouzi D., Zhurkin V.B. Dynamics of Chromatin Fibers: Comparison of Monte Carlo Simulations with Force Spectroscopy // Biophys. J. Elsevier, 2018. Vol. 115, № 9. P. 1644–1655.
40. Ohno M. et al. Sub-nucleosomal Genome Structure Reveals Distinct Nucleosome Folding Motifs // Cell. 2019. Vol. 176, № 3. P. 520-534.e25.
41. Shaytan A.K. et al. Hydroxyl-radical footprinting combined with molecular modeling identifies unique features of DNA conformation and nucleosome positioning // Nucleic Acids Res. Oxford University Press, 2017. Vol. 45, № 16. P. 9229–9243.
42. Chen K. et al. DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing // Genome Res. 2013. Vol. 23, № 2. P. 341–351.
43. Li H., Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform // Bioinformatics. 2009. Vol. 25, № 14. P. 1754–1760.
44. Nordin M. et al. Epigenetic regulation of the Igf2/H19 gene cluster // Cell Prolif. 2014. Vol. 47, № 3. P. 189–199.
45. Oluwadare O., Highsmith M., Cheng J. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data // Biol. Proced. Online. 2019. Vol. 21, № 1. P. 7.

# 9. ПРИЛОЖЕНИЕ

Таблица 1. Использованные в работе данные

| Номер набора данных | Тип эксперимента | Клеточная линия | Репозиторий | Код доступа |
|---|---|---|---|---|
| 1 | Micro-C | H1-hESC | 4DNucleome | 4DNFIPKVL9YI |
| 2 | Micro-C | H1-hESC | 4DNucleome | 4DNFI2YHYAJO |
| 3 | Micro-C | HeLa-S3 | 4DNucleome | 4DNFI8HVQBVR |
| 4 | Micro-C | HeLa-S3 | 4DNucleome | 4DNFI9UM7XYK |
| 5 | Micro-C | JM8.N4 (mESC) | 4DNucleome | 4DNFI8HJHIYX |
| 6 | Micro-C | JM8.N4 (mESC) | 4DNucleome | 4DNFI8HJHIYX |
| 7 | MNase-seq | H1-hESC | NucPosDB | GSM1194220 |
| 8 | MNase-seq | HeLa | NucPosDB | GSM1602359 |
| 9 | MNase-seq | mESC | NucPosDB | GSM2183911 |