Тема: "Поиск и описание хромосомных перестроек в хромосоме 1A большой синицы *(Parus major)*"
Данная работа является выпускной квалификационной работой.

Состав коллектива:

- Фишман Вениамин Семёнович, к. б. н., в. н. с., зав. сектора геномных механизмов онтогенеза ИЦИГ СО РАН, руководитель
- Торгашева Анна Александровна, к. б. н., с. н. с., зав. лаборатории рекомбинационного и сегрегационного анализа, руководитель
- Козырева Светлана Юрьевна, лаборант ИЦИГ СО РАН, студентка ФЕН НГУ, 4 курс, группа 19410, исполнитель

Аннотация:

С использованием данных Hi-C-секвенирования синиц с комплексной хромосомной перестройкой и синиц дикого типа, были получены Hi-C-карты контактов хроматина, с помощью которых были с определены границы инверсии в перестройке с точностью до 1000 п. о. Был разработан алгоритм поиска дополнительных последовательностей, отсутствующих в референсном геноме, с использованием данных секвенирования для исследуемого образца и дикого типа. С использованием предложенного алгоритма обнаружены отсутствующие в референсном геноме синицы высокоповторённые последовательности ДНК, суммарная длина которых составляет около 3 миллионов пар оснований в геноме животных дикого типа и около 14 миллионов пар оснований в геноме животных с комплексной хромосомной перестройкой.

Научное содержание работы:
1. Постановка задачи:
Задачи работы можно разделить на биологическую и техническую части. Биологическая цель работы – охарактеризовать перестройки в хромосоме 1A большой синицы *(Parus major)*. Комплексная хромосомная перестройка включает в себя инверсию и районы с вариациями числа копий. На данный момент стандартные методы исследования хромосомных перестроек не позволяют надёжно определить границы и структуру такой комплексной хромосомной перестройки. Поэтому для этого использовались данные полногеномного и Hi-C-секвенирования, которое хорошо подходит для описания больших (в масштабе хромосом) структурных перестроек, таких как инверсия, обнаруженная в хромосоме 1A большой синицы. Метод Hi-C также может определить границы этой инверсии, несмотря на ассоциированные с ней комплексы CNV, и локализовать возможные инсерции последовательностей в гаплотипе с комплексной хромосомной перестройкой.
Техническая цель – используя уже полученные данные NGS секвенирования (не прибегая к дорогостоящим методам секвенирования третьего поколения), найти предполагаемые высокоповторённые последовательности в гомологе с комплексной перестройкой, которые отсутствуют в референсном геноме. В настоящее время не существует способов найти такие последовательности в данных NGS секвенирования. Для этого был разработан алгоритм поиска дополнительных последовательностей, отсутствующих в референсном геноме, с использованием данных секвенирования для исследуемого образца и дикого типа.

2. Современное состояние проблемы:

Ранее в популяции на территории Нидерландов [1] с помощью генотипирования более 2300 особей и секвенирования геномов более 20 птиц был выявлен полиморфизм по комплексной перестройке в хромосоме 1A большой синицы (Parus major). Согласно данным анализа гетерозиготности и распределения количества однонуклеотидных полиморфизмов (Single Nucleotide Polymorphism, SNP), перестройка предположительно включает в себя большую (более 90% длины хромосомы) инверсию. Один из ее концов, вероятно, находится вблизи комплекса, для которого характерна вариация числа копий (copy number variations, CNV) [1]. При этом у 96% гетерозигот по комплексной перестройке в хромосоме 1A число копий этого комплекса было существенно выше, чем у гомозигот по нормальной хромосоме 1A. Общая длина последовательностей, дополнительно присутствующих в перестроенной хромосоме, предположительно могла составлять около 3.5 Мб [1], что затрудняет определение точных границ инверсии.

С помощью методов цитогенетического анализа наши коллеги из Института Цитологии и Генетики обнаружили полиморфизм, предположительно, по той же хромосомной перестройке в хромосоме 1A, у синиц из новосибирской популяции. Однако длина перестроенной хромосомы была примерно в 1,5 раза больше, чем нормальной, т.е. увеличение длины гомолога, несущего инвертированный аллель, составляло до 30 Мб – существенно больше, чем увеличение на 3.5 Мб, описанное раньше. Кроме того, конфигурации бивалентов в мейозе, в том числе инверсионных петель, косвенно указывали на то, что инверсия, вероятно, не включает центромеру и короткое плечо, на котором расположены высококопийные последовательности. Таким образом, структура этой перестройки – точные границы инверсии, а также генетическое содержание инсерции, неизвестны.

The Genomic Complexity of a Large Inversion in Great Tits / V. H. Da Silva [и др.] // Genome Biology and Evolution. — 2019. — Июль. — Т. 11, № 7. — С. 1870—1881. — ISSN 17596653. — DOI: 10.1093/gbe/evz106. — URL: https://academic.oup.com/gbe/article/11/7/1870/5494702.

3. Подробное описание работы, включая используемые алгоритмы

Согласно цитологическим данным, полученным в лаборатории рекомбинационного и сегрегационного анализа ИЦИГ СО РАН, в популяции большой синицы на территории Новосибирска была обнаружена предположительно та же, что и на территории Нидерландов, комплексная хромосомная перестройка. Однако длина перестроенной хромосомы была примерно в 1,5 больше, чем нормальной (т.е. до 30 Мб больше, а не 3.5 Мб). Кроме того, конфигурации бивалентов в мейозе, в том числе инверсионных петель, косвенно указывали на то, что инверсия, вероятно, не включает центромеру и короткое плечо, на котором расположены высококопийные последовательности.

Для того, чтобы описать этот полиморфизм, была построены Hi-C карты на основе данных секвенирования Hi-C библиотек для синицы с комплексной хромосомной перестройкой и синицы дикого типа. При анализе полученных карт были обнаружены локусы с аномальным распределением числа контактов хроматина, мы предположили, что они скорее всего являются ошибками сборки референсного генома. Для уточнения сборки, данные обработали с помощью программы 3D-DNA, затем вручную переставляли и (или) инвертировали аномальные участки генома.

Для того, чтобы найти дополнительные последовательности в гаплотипе с комплексной хромосомной перестройкой, был разработан алгоритм поиска дополнительных последовательностей, отсутствующих в референсном геноме, с использованием данных

секвенирования для исследуемого образца и дикого типа. Он состоит из следующих этапов:

1. Выравнивание данных WGS и Hi-C-секвенирования на референсный геном.
2. Анализ данных Hi-C-секвенирования методом скользящего окна. Получить сегменты выравнивания, у которых парное прочтение не картируется на геном для каждого участка с шагом в половину длины окна. В данной работе длина окна составляет 200 000 п.о., а шаг 100 000 п.о. Для каждого участка получить последовательности некартируемых сегментов выравнивания и составить из них последовательность.
3. Получить частоты k-меров в ней с помощью инструмента Jellyfish для каждого участка.
4. Анализ k-меров для участков, найденных на предыдущем шаге. Попарно сравнить частоты k-меров для каждого участка в исследуемом образце и образце дикого типа. В данной работе k = 80 п.о. Если в участке в исследуемом образце есть искомая последовательность, ожидается что количество некоторых k-меров в нём будет заметно больше, чем в диком типе. Для удобства поиска таких k-меров можно визуализировать эти данные с помощью графиков, где для каждого k-мера на одной оси отложено количество его повторов в исследуемом образце, а на другой оси – в диком типе. Если участков много, можно использовать метрики, которые позволяют отфильтровать регионы, в которых частоты k-меров для исследуемого образца мало отличаются от частот в диком типе. В данной работе в качестве такой метрики использовалась линейная регрессия. После фильтрации для каждого участка нужно построить график. Ожидается найти k-меры, частоты которых в исследуемом образце будут сильно отличаться от частот в диком типе. Эти k-меры предположительно составляют искомые последовательности.
5. Проанализировать последовательности найденных k-меров. Убедиться, что последовательности найденных k-меров отсутствуют в референсном геноме с помощью сервиса nucleotide BLAST или с помощью BWA-MEM. После этого из k-меров составить одну или несколько консенсусных последовательностей.
6. Добавить полученные последовательности в референсный геном.
7. Произвести выравнивание данных WGS и Hi-C-секвенирования на геном с добавленными последовательностями и проанализировать полученные данные. Проанализировать выровненные на добавленные последовательности прочтения, по возможности увеличить длину этих последовательностей с обоих концов. Для этого нужно посмотреть на структуру сегментов выравнивания, которые фланкируют добавленные последовательности. Если в данных полногеномного секвенирования будет большое количество прочтений, части которых «выходят за пределы» добавленных последовательностей, и они будут составлять одну последовательность, то можно достроить добавленную в геном последовательность. После этого при необходимости повторить выравнивание данных секвенирования.
8. Оценить копийность найденных последовательностей и локализовать добавленные последовательности с помощью данных Hi-C-секвенирования. Построить Hi-C-карту для изменённой версии генома. Если добавленная последовательность имеет контакты с нормальным распределением, то можно предположить, что найденная последовательность действительно присутствует в последовательности исследуемого генома. Если покрытие этой

последовательности будет не ниже, чем среднее по геному, то это также косвенно подтверждает эту гипотезу.

9. Повторить шаги 3-8 для генома с добавленной последовательностью.

4. Полученные результаты:

1. С помощью метода Hi-C уточнена сборка хромосомы 1A большой синицы и определены координаты инверсионного полиморфизма, описанного ранее для этой хромосомы, с точностью до 1000 пар оснований.
2. Впервые дана оценка длины районов с вариациями числа копий на хромосоме 1A на основе полногеномного секвенирования. Суммарная длина повторённых участков на гомологе хромосомы 1A, несущем инверсионный полиморфизм, составляет 4-5 миллионов пар оснований.
3. Предложен алгоритм для поиска высокоповторённых последовательностей, отсутствующих в референсном геноме, на основе данных Hi-C-секвенирования. С использованием предложенного алгоритма обнаружены отсутствующие в референсном геноме синицы высокоповторённые последовательности ДНК, суммарная длина которых составляет около 3 миллионов пар оснований в геноме животных дикого типа и около 14 миллионов пар оснований в геноме животных с комплексной хромосомной перестройкой. Таким образом, объяснено увеличение гомолога хромосомы 1A с комплексной хромосомной перестройкой на 11 миллионов пар оснований по сравнению с гомологом дикого типа.

5. Иллюстрации, визуализация результатов:

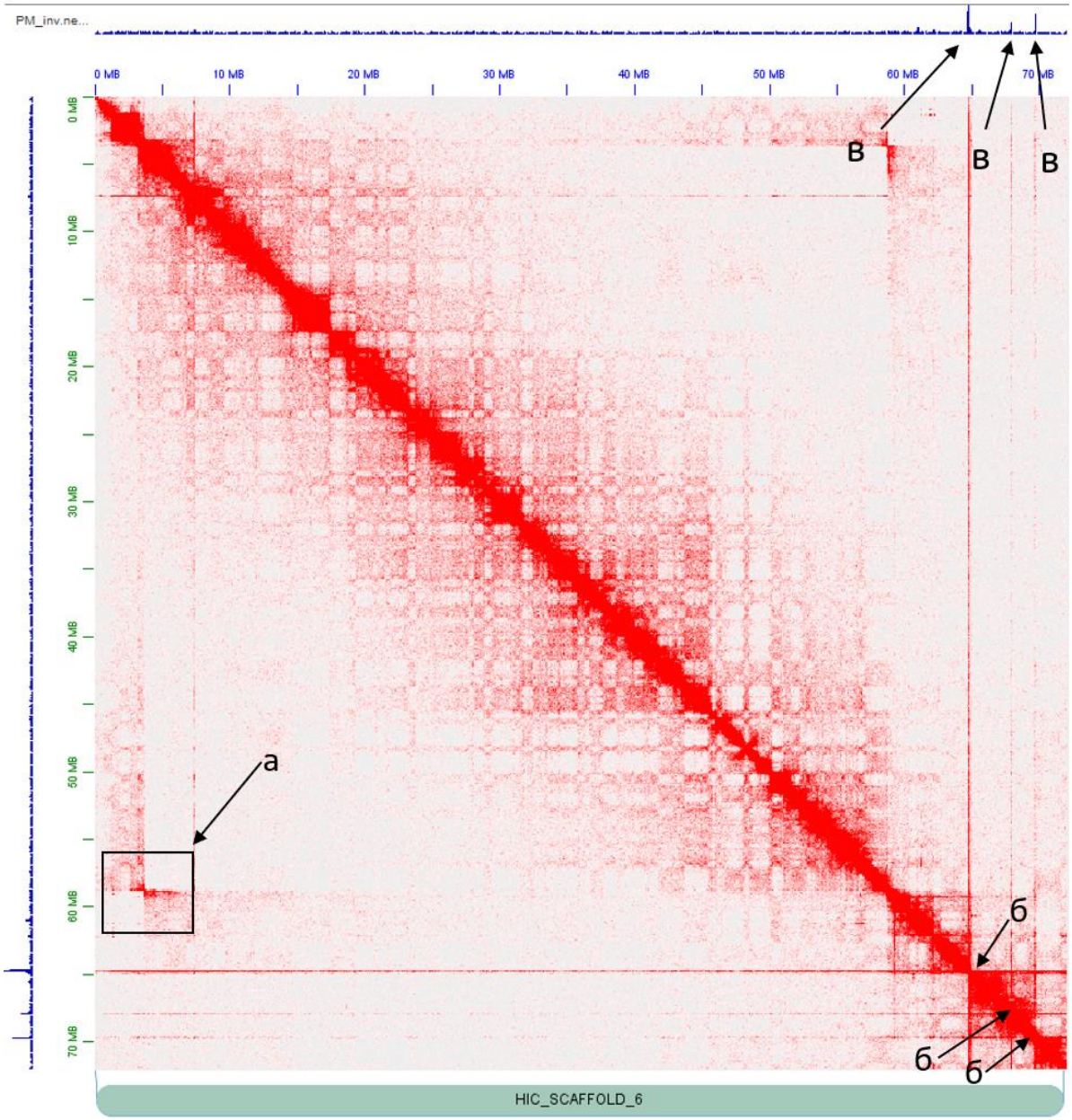

Рисунок 1. Hi-C карта хромосомы 1A большой синицы с комплексной перестройкой после обработки программой 3D-DNA и Juicebox Assembly.
На карте хорошо виден характерный рисунок в виде бабочки, свидетельствующий о наличии инверсии (а). Также на карте видны районы с вариациями числа копий (б). Это также подтверждается графиком покрытия сегментами выравнивания, расположенном вверху и слева от карты (в).

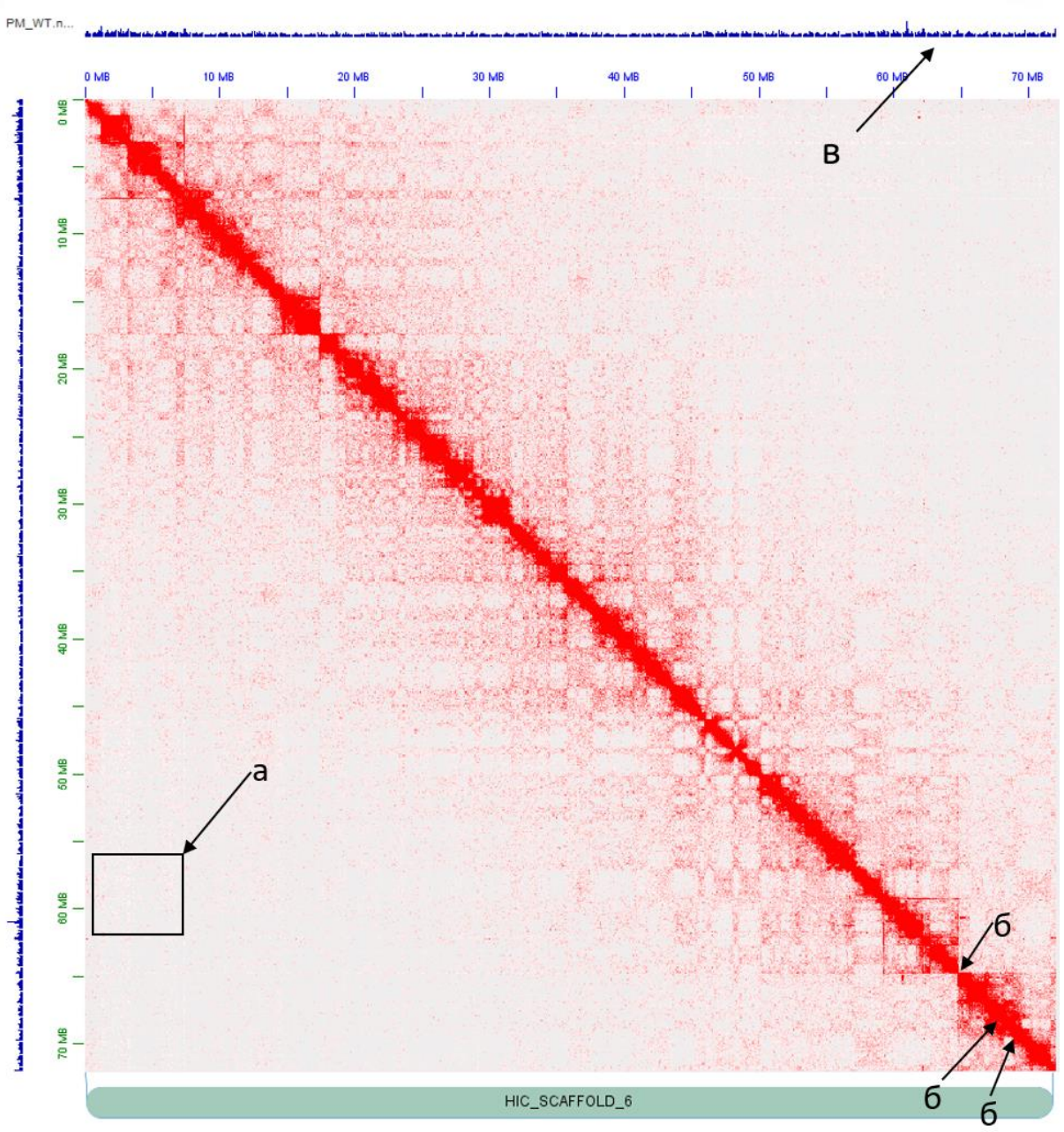

Рисунок 11. Hi-C карта хромосомы 1A дикого типа большой синицы.
На карте отсутствует характерный для инверсии паттерн контактов рисунок в виде бабочки (а). Также на карте не наблюдаются высококопийные районы CNV (б), которые видны на аналогичной карте для хромосомы с комплексной перестройкой. Это также подтверждается графиком покрытия сегментами выравнивая, расположенным вверху и слева от карты (в).

Эффект от использования кластера в достижении целей работы.
Данные NGS секвенирования занимают несколько десятков гигабайт дискового пространства. Выравнивание данных таких объёмов на референсный геном невозможно провести на локальной машине из-за большого объёма оперативной памяти и времени вычислений. Больших вычислительных ресурсов также требует получение Hi-C-карт и их обработка. Выравнивание данных секвенирования и построение Hi-C-карт в данной

работе проводилось большое количество раз, поэтому без использования кластера невозможно достичь целей работы.