

Отчёт о проделанной работе с использованием оборудования ИВЦ НГУ

1. Аннотация.

Была проведена сборка и аннотация геномов полёвок видов *Microtus arvalis* и *Alexandromys oeconomus*. Качество полученных сборок было проверено по стандартному протоколу программой BUSCO, в качестве референса были выбрана база генов грызунообразных (параметр -l glires). Результаты проверки показывают сохранение более 98% генов для обоих видов, что указывает на высокое качество полученных сборок.

В соответствии с планом работ, были полученные геномы были аннотированы с помощью программы RepeatMasker версии 4.1.7-p1, входящей в конвейер программ RepeatModeller, по базе повторов Dfam 3.8.1 В обоих видах ~37% генома покрывают различные повторяющиеся элементы, из которых наибольшую долю составляют SINE и LTR (~14%). Межвидовые различия в представленности разных групп повторов минорны. В результате поиска эволюционно консервативных элементов в геномах полёвок исследуемых видов и *M. musculus* была выявлено 400-430 синтенных блоков со средней длиной ~4.5-5.5 млн п.о. и медианной длиной ~1.3 млн. п.о. При сравнении *Mc. arvalis* и *A. oeconomus* выделено ~200-220 блоков синтении средней длиной около 9-10 млн. п.о. и медианной длиной от 3 до 4.5 млн. п.о. Данные различия отражают эволюционную близость *Mc. arvalis* и *A. oeconomus* и их общую равную удалённость от *M. musculus*. Далее был проведен анализ сходства профилей экспрессии между разными особями одного вида на разрешении 1 тысяча пар оснований. Для всех трёх особей *Mc. arvalis* коэффициент корреляции экспрессии составил от 0.88 до 0.99 ($p < 0.05$), что указывает на небольшую изменчивость между особями в профиле экспрессии генов и возможность их совместного использования в эволюционном анализе. В тоже время, для *A. oeconomus* коэффициент корреляции между двумя биологическими репликами составил всего 0.61. ($p < 0.05$). Таким образом эволюционной сравнение взаимосвязи профилей экспрессии и пространственной организации хроматина оказывается под большим влиянием изменчивости между особями *A. oeconomus*.

Результаты РНК-секвенирования были объединены и проанализированы с помощью программы StringTie (версии 2.2.3), что позволило аннотировать более 20 тысяч транскриптов печени полёвок видов *A. oeconomus* (22 804) и *Mc. arvalis* (26 545).

2. Тема работы:

Сборка генома не модельных организмов до хромосомного уровня

3. Состав коллектива Нуридинов Мирослав Абдурахимович, канд. биол. наук м.н.с. ИциГ СО РАН, nuriddinov@bionet.nsc.ru

4. Информация о грантах

РНФ № 24-24-00321, Эволюция пространственной организации хроматина на субТАДовом уровне у млекопитающих

5. Научное содержание работы.

5.1. Постановка задачи

Сборка и аннотация геномов полёвок видов *Microtus arvalis* и *Alexandromys oeconomus*

5.2. Современное состояние проблемы.

5.3. Подробное описание работы, включая используемые алгоритмы.

На первом этапе с помощью данных Hi-C проводилось улучшение геномных сборок *A. oeconomus* (Aoes_v1, GCA_963919705.1) и *Mc. arvalis* (Marv_v2, GCA_963919715.1) программы 3D-DNA. Сначала сборка проводилась в автоматическом режиме для определения предполагаемых хромосом. Затем в ручном режиме, проводилось исправление имеющихся ошибок сборки. Разрешение предполагаемых гаплотипов проводилось с помощью взаимного выравнивания целевых скаффолдов программой LastZ. Локусы, для которых обнаруживался высокий уровень сходства нуклеотидной последовательности (>50%) и взаимного покрытия

последовательности (>70%) расценивались как гаплотипы. Из нескольких гаплотипов в итоговую сборку отбирались те, которые имели наибольшую длину, наибольшее покрытие прочтениями Hi-C и предполагали наименьшее количество шагов по редактированию последовательности генома. Был проведён поиск контигов и скаффолдов соответствующих Y-хромосоме для *A. oesonotus* и *Ms. arvalis*. Для этого было проведено выравнивание последовательностей, длиной более 15 тысяч пар оснований и имеющих покрытие более 10 пар прочтений на референсный геном *Mus musculus* с помощью программы lastz (версии v1.04.22) со следующими параметрами: --hspthresh=6000 --gfextend --chain --gapped . Из полученных выравниваний предполагалось отобрать те, которые имеют больше 50% длины которых имеют соответствии на Y-хромосоме *M. musculus*, и с помощью программы 3DNA затем собрать в ручном режиме. В ходе улучшения сборки *A. oesonotus*, среди скаффолдов и контигов имеющихся в референсной базе данных удалось выявить те из них, которые достоверно принадлежат Y-хромосоме. Суммарный размер отобранных последовательностей составил 2 129 167 пар оснований, что значительно меньше ожидаемого, исходя из эволюционной консервативности (~90 миллионов пар оснований для *Mus musculus*). Несмотря на явную неполноту Y-хромосомы, она была добавлена к имеющемуся геному *A. oesonotus*. Качество полученных сборок было проверено по стандартному протоколу программой BUSCO, в качестве референса были выбрана база генов грызунообразных (параметр -l glires). Результаты проверки показывают сохранение более 98% генов для обоих видов, что указывает на высокое качество полученных сборок.

Полученные геномы были аннотированы с помощью программы RepeatMasker версии 4.1.7-r1, входящей в конвейер программ RepeatModeller, по базе повторов Dfam 3.8.1. Для проведения сравнительного анализа геномов, выделения консервативных элементов и блоков синтении, с помощью программы LastZ было проведено попарное выравнивание геномов *M. musculus* и *Ms. arvalis* и *A. oesonotus*. Консервативным элементом признавалась последовательность длиной не менее 50 нуклеотидов (--hspthresh=5000), с совпадением не менее 50% нуклеотидов у сравниваемых видов (--filter=identity:50), и невыравниваемые участки не могут занимать более 30% локуса (--filter=continuity:70). Были проанализированы 5 библиотек РНК содержащих 40-44 млн прочтений на образец, что обеспечивает порядка 30-кратного покрытия кодирующей части генома исследуемых видов. Данные библиотеки были выравнены с помощью программы hisat2 (версии 2.2.1) на полученные в ходе этой работы сборки геномов *A. oesonotus* и *Ms. arvalis*.

5.4. Полученные результаты.

В результате были получены сборки генома хромосомного уровня условно названные alOe_r10a для *A. oesonotus* и miAr_r10a для *Ms. arvalis*. Сборка alOe_r10a характеризуется набором хромосом из 14 соматических хромосом и хромосомы X, а сборка miAr_r10a — 22 соматических хромосомы и хромосома X, что соответствует имеющимся цитогенетическим данным.

Аннотация повторов показала, что в обоих видах ~37% генома покрывают различные повторяющиеся элементы, из которых наибольшую долю составляют SINE и LTR (~14%). Межвидовые различия в представленности разных групп повторов минорны. Полученные результаты демонстрируют существенное, почти двукратное уменьшение (с ~13% до 7%) доли LINE повторов в геноме исследуемых видов по сравнению с *M. musculus*. В результате поиска эволюционно консервативных элементов в геномах полёвок исследуемых видов и *M. musculus* была выявлено 400-430 синтенных блоков со средней длиной ~4.5-5.5 млн п.о. и медианной длиной ~1.3 млн. п.о. При сравнении *Ms. arvalis* и *A. oesonotus* выделено ~200-220 блоков синтении средней длиной около 9-10 млн. п.о. и медианной длиной от 3 до 4.5 млн. п.о. Данные различия отражают эволюционную близость *Ms. arvalis* и *A. oesonotus* и их общую равную удалённость от *M. musculus*.

Анализ сходства профилей экспрессии между разными особями одного вида на разрешении 1 тысяча пар оснований. Для всех трёх особей *Ms. arvalis*. коэффициент корреляции

экспрессии составил от 0.88 до 0.99 ($p < 0.05$), что указывает на небольшую изменчивость между особями в профиле экспрессии генов и возможность их совместного использования в эволюционном анализе. В тоже время, для *A. oeconomus* коэффициент корреляции между двумя биологическими репликами составил всего 0.61. ($p < 0.05$) Таким образом эволюционной сравнение взаимосвязи профилей экспрессии и пространственной организации хроматина оказывается под большим влиянием изменчивости между особями *A. oeconomus*. Далее, результаты РНК-секвенирования были объединены и проанализированы с помощью программы StringTie (версии 2.2.3), что позволило аннотировать более 20 тысяч транскриптов печени полёвок видов *A. oeconomus* (22 804) и *Ms. arvalis* (26 545).

Сборки геномов и аннотация доступны в сети Интернет по адресу https://genedev.bionet.nsc.ru/ftp/by_Project/Microtus/results/

6. Эффект от использования кластера в достижении целей работы. Обработка данных по пространственной организации хроматина требует больших объёмов физической и оперативной памяти, что делает её невозможной на обычных рабочих станциях. Использование ресурсов кластера является определяющим для достижения целей работы.

7. Доступ к результатам работы

Непроцессированные данные доступны в репозитории NCBI по адресу: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1378222>

Сборки геномов, их аннотация и Hi-C карты доступны в сети Интернет по адресу https://genedev.bionet.nsc.ru/ftp/by_Project/Microtus/results/