Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ
Заведующий кафедрой
_____ В. А. Кратасюк
подпись инициалы, фамилия
«____» июня 2018г.

БАКАЛАВРСКАЯ РАБОТА

06.03.01 – Биология

АНАЛИЗ ПОВТОРЯЮЩИХСЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ В ГЕНОМАХ *PORODAEDALEA NIEMELAEI*, *P. CHRYSOLOMA* И *ARMILLARIA BOREALIS*

Руководитель _____ д.ф.-м.н., в.н.с. М. Г. Садовский
подпись, дата должность, ученая степень инициалы, фамилия

Выпускник _____ А. И. Аксёнова
подпись, дата инициалы, фамилия

Красноярск 2018

РЕФЕРАТ

Выпускная квалификационная работа «Анализ повторяющихся последовательностей в геномах *Porodaedalea niemelaei, P. chrysoloma* и *Armillaria borealis»* содержит 35 страниц текстового документа, приложение, 37 использованных источников, рисунков, 9 таблиц

ГЕНОМНЫЕ ПОВТОРЫ, ТРАНСПОЗОНЫ, ГРИБНЫЕ ГЕНОМЫ, НУКЛЕОТИДНЫЕ ПОСЛЕДОВАТЕЛЬНОСТИ, МОБИЛЬНЫЕ ЭЛЕМЕНТЫ, МИКРОСАТЕЛЛИТЫ, LINE, SINE, LTR, non-LTR

Цель исследования – выявление повторяющихся нуклеотидных последовательностей в геномах *Porodaedalea niemelaei*, *P. chrysoloma*, *Armillaria borealis* и их классификация

Объектом исследования являются геномы грибов *Porodaedalea niemelaei* M. Fischer, *Porodaedalea chrysoloma* (Fr.) Fiasson & Niemelä и *Armillaria borealis*.

Получены библиотеки повторов для трёх видов. *Porodaedalea niemelaei* содержит 158 повторов, *P. chrysoloma* – 122, а *Armillaria borealis* – 886.

Содержание

ВВЕДЕНИЕ

Повторяющиеся последовательности – это участки ДНК, включённые в геном, последовательность которых состоит из повторяющихся фрагментов.

Такие последовательности вносят значительный вклад в эволюцию транскриптомов, промоторов и протеомов. Повторы ДНК, которые представляют основную часть большинства геномов, вызывают его нестабильность, часто приводящую к перестройкам хромосом и заболеваниям [1].

У высших эукариот от 40% всего генома составляют повторы, у грибов их меньше (около 10%) и они в основном представлены длинными концевыми повторами (LTRs).

Кроме того, значительная часть генома эукариот представлена нетранслируемыми участками – интронами. У грибов интроны обнаружены в небольшом числе генов [2].

Анализ повторяющихся последовательностей в геномах грибов рода *Porodaedalea* и *Armillaria* был начат с целью поиска новых важных генов и метаболических путей, участвующих в разложении древесины и лигнина.

Род *Porodaedalea* играет важную биологическую и экологическую роль в поддержании минерального баланса в бореальных лесных экосистемах, а также участвует в глобальном углеродном цикле. Воздействие этого рода грибов на растения может увеличиться из-за глобального изменения климата. Эти грибы эффективно разлагают древесину, лигнин и целлюлозу на более простые соединения, которые могут быть использованы в производстве этанола на основе целлюлозы.

Armillaria относится к грибам-некротрофам. Этот гриб колонизирует живые корни, убивает корневую ткань, а затем использует мертвые ткани как источник питания. Тем самым паразит избегает иммунных реакций, которые

присущи всем живым клеткам, а гибель участков зараженного растения опережает заселение их гифами паразита [3].

До сих пор не до конца ясна роль и функциональное значение повторов. Постепенно накапливаются сведения о том, что они могут играть важную роль в эволюционном развитии организмов, процессе репликации и формирования нуклеопротеиновых комплексов, а также влиять на регуляцию генной экспрессии [4]. Маскировка повторов, которая является важным этапом при аннотации геномов, не может быть выполнена без точного и полного обнаружения всех повторяющихся последовательностей.

Объектом исследования являются геномы грибов *Porodaedalea niemelaei* M. Fischer , *Porodaedalea chrysoloma* (Fr.) Fiasson & Niemelä и *Armillaria borealis*.

Предмет исследования: повторяющиеся последовательности в геномах *Porodaedalea niemelaei, P. Chrysoloma* и *Armillaria borealis*.

Целью настоящей работы является нахождение повторяющихся последовательностей в геномах *P. niemelaei*, *P. chrysoloma* и *Armillaria borealis*, их классификация, а также сравнение с ранее изученными родственными видами.

В связи с этим были поставлены следующие задачи:

– обзор имеющихся сведений о функциональном значении повторяющихся последовательностей;

– обзор программного обеспечения для поиска повторяющихся последовательностей;

– поиск повторяющихся нуклеотидных последовательностей в геномах *P. niemelaei*, *P. chrysoloma* и *Armillaria borealis*;

– обзор программного обеспечения для выявления неклассифицированных повторяющихся последовательностей;

– классификация неизвестных последовательностей в геноме;

– анализ полученных результатов.

Исследования повторяющихся последовательностей позволят сформировать более точные представления о структуре генома, также они необходимы для аннотирования геномов, филогенетических и эволюционных исследований.

Выпускная квалификационная работа выполнена в лаборатории лесной геномики СФУ в рамках проекта «Геномные исследования основных бореальных лесообразующих хвойных видов и их наиболее опасных патогенов в Российской Федерации», руководимого проф. К. В. Крутовским и финансируемого Правительством РФ (договор №14.Y26.31.0004). Автор работы выражает искреннюю благодарность И.Н. Павлову за предоставленные образцы грибов, Орешковой Н.В. за пробоподготовку и секвенирование, Путинцевой Ю.А. и Садовскому М.Г. за руководство, а также всем членам лаборатории за участие в обсуждении результатов работы и ценные советы.

1 ОСНОВНАЯ ЧАСТЬ

1.1 Обзор литературы

1.1.1 Характеристика повторяющихся элементов

В настоящее время различают несколько классификаций повторяющихся элементов. В основном, все повторы принято делить на тандемные и диспергированные.

Тандемные повторы – множественные копии последовательностей, следующие друг за другом. Например, ТАТАТАТА – это тандемный повтор динуклеотида ТА, CGGCGGCGG – тандемный повтор тринуклеотида CGG. В зависимости от их размера они делятся на:

– сателлитные нуклеотидные последовательности, содержание которых в геномах эукариот может достигать 5–50% от суммарного количества ДНК, представляют собой очень длинные (в несколько сотен тысяч пар нуклеотидов) участки ДНК с тандемно («голова к хвосту») повторяющимися короткими блоками (5–200 п.н.);

– микросателлитные ДНК, составленные тандемными повторами длиной 1–4 п.н., организованы в блоки до 200 п.н., которые рассеяны по геному. В отличие от сателлитной и минисателлитной ДНК, микросателлиты, как правило, транскрибируются. Длина микросателлитов и их общее количество в геноме коррелирует с размером генома [5];

– минисателлитные ДНК построены из повторяющихся последовательностей длиной в 5–50 п.н., формируют блоки промежуточных размеров (до 104 п.н.) и локализованы в разных участках хромосом. Используются в качестве ДНК-маркеров. Механизмами происхождения являются "проскальзывания" при репликации ДНК, точечные мутации и рекомбинация [6].

Диспергированные повторы рассеяны (диспергированы) по всему геному, это фрагменты ДНК, имеющие специальную структурную организацию, которые обладают способностью перемещаться в геноме как в пределах одной хромосомы, так и между хромосомами.

Различают два основных класса таких повторов: транспозоны или ДНК-транспозоны (I) и ретротранспозоны (II). Такая классификация основана на механизмах, с помощью которых перемещаются подвижные элементы.

Транспозоны – это участки ДНК организмов, способные реплицировать и внедрять одну из копий в новое место генома, обычно имеющие длину 2500-7000 нуклеотидных пар.

Большая часть эукариотических геномов состоит из последовательностей, которые возникли, прямо или косвенно, в результате деятельности транспозонов. Большинство этих геномных элементов считаются несущественными для выживания. Однако они оказывают значительное влияние на эволюцию генома [7]. Транспозоны ограничены с двух сторон инвертированными повторами, то есть последовательностями, направленными навстречу друг другу. Инвертированные повторы сближаются и точно отрезаются от соседних участков ДНК хозяина. Вырезанный транспозон внедряется в район вносимого транспозазой разрыва в молекуле-мишени и сшивается с ДНК хозяина в новом месте. Подвижность элементов становится возможной благодаря активности ферментов, которые способны точно вырезать элемент из хромосомы для того, чтобы затем вставить его в другое место генома.

ДНК-транспозоны типичны для генома бактерий и достаточно широко представлены в геномах эукариот. Их транспозиция осуществляется, как правило, по механизму вырезания и вставки (cut and paste) с участием транспозазы – хорошо изученного фермента класса рекомбиназ.

Подкласс 1 представлен порядком TIR (Terminal Inverted Repeats), характерной особенностью которого является наличие концевых инвертированных повторов – TIR – на обоих концах элемента (рис.1).

Транспозиция элементов первого подкласса происходит с помощью фермента транспозазы.

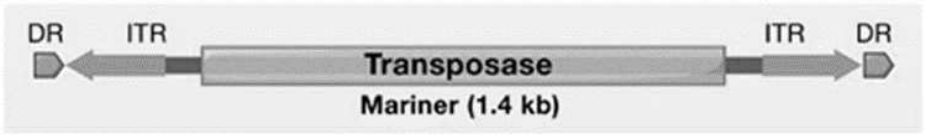

Рисунок 1 – Структура ДНК-транспозона

Подкласс 1 представлен надсемействами Tc-Mariner, hAT, Mutator (MULU), P, PIF-Harbinger и CACTA [8] Наиболее подробно изученные hAT-семейства – Ac-Ds-элементы кукурузы и Tam3 львиного зева.

Подкласс 2 в геномах растений представлен элементами порядка Helitron, которые хорошо описаны в геноме кукурузы [9].

Ретроэлементы широко распространены у растений, где они часто являются важным компонентом ядерной ДНК. У кукурузы 49-78 % генома состоит из ретротранспозонов [10], у пшеницы около 90 % генома представлены повторяющимися последовательностями, из них 68 % – перемещающимися элементами [11]. У млекопитающих, практически половина генома (45-48 %) состоит из транспозонов или остатков транспозонов. Примерно 42 % генома человека состоит из ретротранспозонов, и около 2-3 % из ДНК-транспозонов.

Ретроэлементы, в отличие от ДНК-транспозонов, используют механизмы, в которых важную роль играет обратная транскрипция – синтез ДНК на матрице РНК обратной транскриптазой.

На основании особенностей структуры и механизмов репликации ретроэлементы, в свою очередь, разделяют на большие группы:

1) LTR-содержащие ретроэлементы, включающие ретротранспозоны и эндогенные ретровирусы (long terminal repeat – LTR) [12].

Полноразмерный автономный LTR-ретротранспозон имеет размер от 4 до 10 тыс. пар нуклеотидов и содержит в своей структуре (рис. 2) длинные

концевые повторы в прямой ориентации, длина которых варьирует от 100 п.н. до 5 тыс. п.н. Длинные концевые повторы не кодируют белки, но содержат промоторы и терминаторы, регулирующие транскрипцию генов LTR-ретротранспозонов. LTR-ретротранспозоны могут содержать дополнительные кодирующие последовательности. Например, некоторые ретротранспозоны растений содержат открытые рамки считывания, кодирующие белки ретровирусов. Роль таких белков остается пока невыясненной [13].

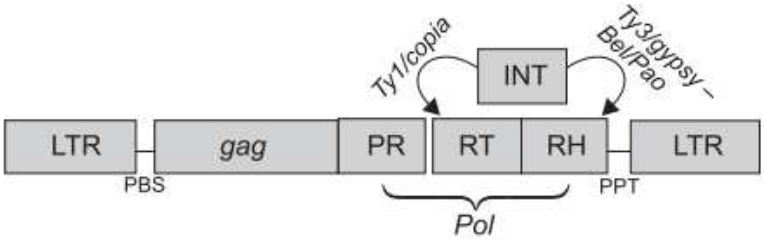

Рисунок 2 – Структура LTR-ретротранспозона

Главным отличием ретровирусов от LTR-ретротранспозонов группы Ty3/gypsy является то, что ретровирусы имеют дополнительную открытую рамку считывания (ORF) для кодирования белков оболочки, необходимых для передачи ретровируса от клетки к клетке. Ty3/gypsy LTR-ретротранспозоны имеют длину от 4 до ~ 15 тыс. п.н., длинные концевые повторы LTR, открытую рамку считывания, в которой закодированы гены *gag* и *pol* (в случае ретровирусов еще и ген env), и некоторые дополнительные гены (в случае Tat элементов и различных хромовирусов) [14].

Суперсемейство Ty1/copia. Ty1/copia представляет собой суперсемейство ретровирусов и LTR-ретротранспозонов, широко представленных в геномах растений, грибов, животных, водорослей и некоторых простейших. Среди представителей суперсемейства Ty1/copia описаны Sire-элементы из геномов растений, содержащие дополнительную ORF, кодирующую ген *env*, благодаря чему им был присвоен статус потенциальных ретровирусов. Ty1/copia LTR-ретротранспозоны имеют LTRs длиной 100–1300 п.н., ограничивающие

центральный район с одной открытой рамкой считывания, в которой закодированы гены *gag* и *pol* (в случае ретровирусов еще и ген *env*) [15].

2) nonLTR-ретротранспозоны (рис. 3), не несущие длинных концевых повторов (рис. 3). Вторую группу также называют LINE-элементами (Long Interspersed Nuclear Elements – длинные диспергированные ядерные элементы);

Рисунок 3 – Структура nonLTR-ретротранспозона

3) Penelope–like элементы (PLEs), кодирующие обратную транскриптазу (RT), более близкую к теломеразе, чем к обратной транскриптазе LTR-ретротранспозонов, и эндонуклеазу, более близкую к эндонуклеазе интронов группы II и бактериальному белку UvrC (рис. 4);

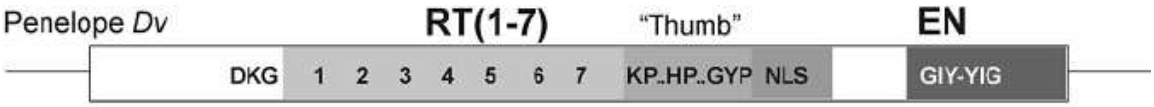

Рисунок 4 – Структура Penelope-like элемента

4) SINE-элементы (Short Interspersed Nuclear Elements) – короткие диспергированные ядерные элементы, не кодируют белки (рис. 5) [16].

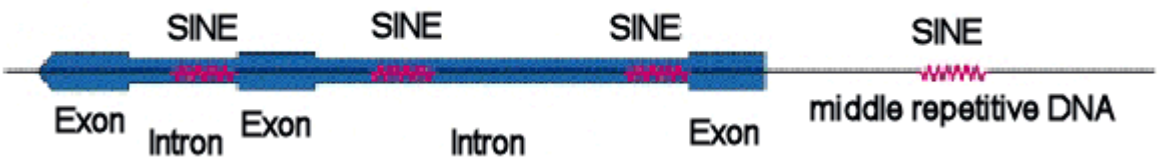

Рисунок 5 – Структура SINE-элемента

1.1.2 Обзор повторяющихся последовательностей в геномах грибов

Последние 10 лет были продуктивными в изучении транспонируемых элементов (TEs) в грибах. Все описанные эукариотические TEs найдены, включая распространенных активных членов семейства *Pogo*.

Аско- и базидиомицеты редко содержат более 5% TEs, тогда как для филогенетически более старого отдела грибов - *Zygomycota* характерно более 30% TEs.

Причина, по которой грибы предпочитают использовать как модельные системы для молекулярного анализа – их сравнительно малый размер генома, который, как правило, содержит только ограниченное количество повторяющихся последовательностей. *Aspergillus nidulans* с размером генома в 27 Mb, имеет лишь 5% повторяющихся последовательностей [17]. В других хорошо изученных видах (*Neurospora crassa* и *Saccharomyces cerevisiae*) ситуация аналогична. *N. crassa* имеет геном размером в 27 Mb и содержит не более 8% повторяющихся последовательностей [18].

Ситуация отличается для более древнего отдела грибов – *Zygomycota*.

Зигомицеты имеют размер генома в пределах между 42 Mb (*Phycomyces blakesleeanus, Phycomycetaceae*) [19] и 54 Mb (*Absidia glauca, Absidiaceae*) [20]. В этих зигомицетах процент TEs составляет около 35%.

Роль TEs в мутации и организации генома хорошо документирована, что приводит к существенному прогрессу в нашем восприятии механизмов, лежащих в основе генетических изменений в этих организмах. TE-опосредованные изменения, связанные с транспозицией и рекомбинацией, обеспечивают широкий диапазон генетических вариаций, который полезен для естественных популяций в их адаптации к экологическим ограничениям, особенно для тех, у кого отсутствует половая стадия. Интересно, что некоторые виды грибов развили различные механизмы замалчивания, которые

рассматриваются как системы защиты хозяев от TEs. Изучение сил, воздействующих на эволюционную динамику ТЕ, должно обеспечить важную информацию о взаимодействиях между ТЕ и геномом грибов [21]. Более того, ТЕ могут играть другие роли в эволюции, связанные с их способностью переноситься горизонтально, захватывать и переносить хромосомные последовательности хозяев, тем самым обеспечивая механизм для диспергирования последовательностей на новых сайтах. Однако, активность переносимых элементов и, следовательно, их пролиферация в геноме хозяина может быть затронута некоторыми видами грибов, которые подвергаются мейозу, путем замораживания процессов. Наше понимание биологических эффектов TEs на грибковый геном резко возросло за последние несколько лет, и выяснение того, в какой степени транспозоны способствуют генетическим изменениям в природе, обеспечивая гибкость адаптации организмов к изменениям окружающей среды, является важной областью для будущих исследований.

1.1.3 Обзор подходов к поиску повторяющихся нуклеотидных последовательностей

Существуют различные категории программ в соответствии с их методологией и характером анализируемых повторов, которые они могут идентифицировать.

Выбор метода для поиска повторяющихся последовательностей зависит от уровня знаний о типах повторов, встречающихся в исследуемом геноме. Можно провести поиск конкретных типов повторов (например, поиск только коротких тандемных повторов, поиск ретротранспозонов), поиск повторов, имеющих характерные структурные особенности, и поиск совершенно новых и неизвестных элементов только на основе их повторяющегося характера [22].

Поиск повторяющихся последовательностей с использованием библиотек повторов. Этот метод основан на сравнении входных данных (например, собранного генома) с референсными последовательностями, содержащимися в библиотеке.

Это может быть как библиотека, собранная вручную (с учетом всех требований и поставленных задач), так и классическая библиотека REPBASE [23]. Эта библиотека содержит консенсусные последовательности повторов различных эукариотических организмов. Наиболее распространенной программой для такого поиска повторов с использованием библиотек является RepeatMasker. Программа была первоначально разработана для маскировки повторов, чтобы облегчить дальнейшее исследование генома. В случае короткого повтора, когда внутренние делеции - вставки маловероятны, для поиска используется простое сканирование последовательности заданной моделью. Выходные данные содержат подробную аннотацию обнаруженных повторов, а также модифицированную версию входных последовательностей, в которых повторы были заменены на «N».

Другие (менее востребованные) программы, использующие аналогичный подход: Censor 24 , Maskeraid, который предназначен для повышения производительности RepeatMasker, PLOTREP, и Greedier.

Основной недостаток программ, основанных на поиске сходства с ранее составленными библиотеками, кроется в самом принципе: полностью основанный на гомологии, этот принцип позволяет обнаружить только уже известные повторяющиеся последовательности, и абсолютно нечувствителен к новым элементам. Тем не менее, программы типа RepeatMasker часто используются в качестве первого этапа *ab initio* идентификации повторяющихся элементов, так как они способны генерировать новые библиотеки повторов.

Сигнатурный подход (The signature-based approaches). Этот подход может быть использован, чтобы найти новые элементы, но не новые классы элементов. Использование таких методов целиком и полностью зависит от того, насколько хорошо известна структура элементов, принадлежащих к определенным

классам. Некоторые классы элементов более структурированы, чем другие, и это приводит к смещению в сторону обнаружения классов с ярко выраженными структурными характеристиками; метод менее чувствителен к классам повторов с менее выраженной структурой.

De novo подходы. Позволяют проводить поиск повторов любого рода без опоры на библиотеки или ранее известные структурные характеристики. Эти подходы предназначены в основном для обнаружения новых повторов и для работы с геномами, по которым накоплено недостаточно информации. Существуют два основных подхода к обнаружению повторяющихся последовательностей. Первый состоит в сравнении последовательностей генома друг с другом, а второй - на поиске повторяющихся *k*-меров (последовательность из *k*-нуклеотидов).

Идентификация семейств повторов. Некоторые из вышеупомянутых программ выполняют функцию кластеризации повторов по семействам (алгоритм «сверху – вниз»). Но в основном программы сначала обнаруживают повторы, а затем предлагают один из способов определения семейства (алгоритм «снизу – вверх»).

Программа RepeatGluer [25] основана на алгоритме графов де Брейна (рис. 6). Графы представляют каждый *k*-мер в геномной последовательности в качестве узла. Затем два узла соединяются направленным ребром, если они накладываются друг на друга в геноме. Строится консенсусная последовательность внутри каждого семейства, и определяется количество повторений.

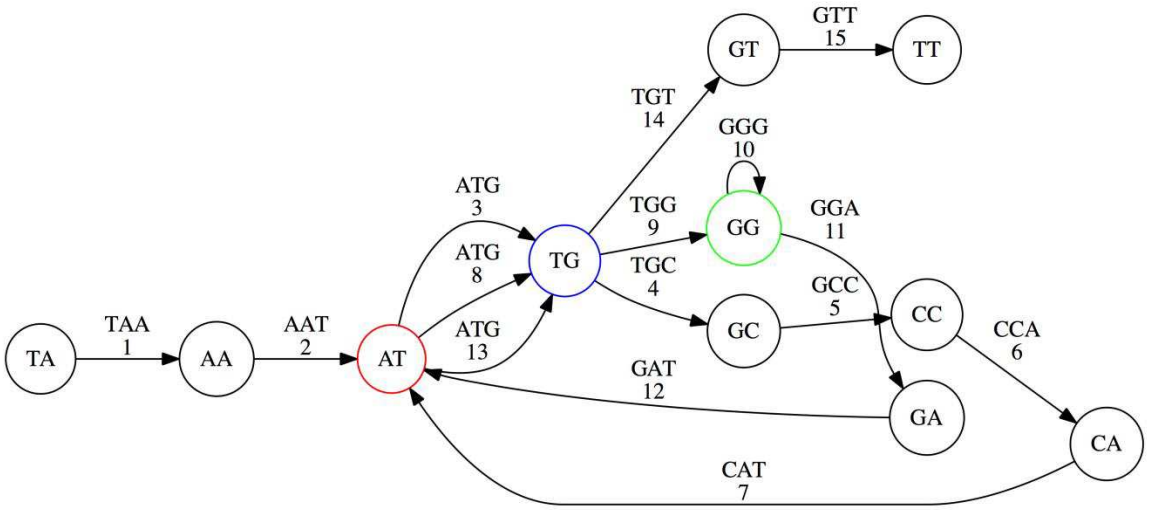

Рисунок 6 – Пример графа де Брёйна

Сложность определения различных типов повторяющихся фрагментов тесно связана с мутационными процессами, происходящими в организме, благодаря которым происходят вставки, замены и делеции отдельных нуклеотидов, а иногда и целых участков ДНК. Большинство подходов к анализу нуклеотидных и аминокислотных последовательностей основано на алгоритмах работы с текстовыми строками; в этом случае учет точечных мутаций является тяжелой с вычислительной точки зрения операцией, поскольку он вносит существенную нелинейность в подобные алгоритмы [26].

2 МАТЕРИАЛЫ И МЕТОДЫ

2.1 Методы исследования

Геном *Porodaedalea niemelaei* секвенировали и аннотировали с использованием технологии PacBio в JGI (Joint Genome Institute) совместно с лабораторей лесной геномики СФУ. Размер полногеномной сборки составил 53,34 млн. нуклеотидных оснований, итоговая сборка генома состоит из 951 скаффолда, показатель геномной сборки N50 составил 140 тыс. п.н (таблица 1).

Таблица 1 – Характеристики геномных сборок *P. niemelaei*

Показатель	***P. niemelaei***
Количество контигов	951
Размер сборки (в млн. п. н.)	53,34
N50 по скаффолдам (тыс. п. н.)	140
Место взятия образца	Россия(Таймыр)

Геном *Porodaedalea chrysoloma* был взят из базы данных JGI. Размер полногеномной сборки составил 44,69 млн. нуклеотидных оснований, итоговая сборка генома состоит из 1684 скаффолдов, показатель геномной сборки N50 составил 550 тыс. п.н. (таблица 2).

Таблица 2 – Характеристики геномных сборок *P. chrysoloma*

Показатель	***P. chrysoloma***
Количество контигов	1684
Размер сборки (в млн. п. н.)	44.69
N50 по скаффолдам (тыс. п. н.)	550
Место взятия образца	США

Геном *Armillaria borealis* был секвенирован и аннотирован в лаборатории лесной геномики СФУ под руководством проф. К. В. Крутовского. Размер полногеномной сборки составил 66,79 млн. нуклеотидных оснований, итоговая сборка генома состоит из 46 249 скаффолдов, показатель геномной сборки N50 составил 649 тыс. п. н.

Таблица 3 – Характеристики геномных сборок *A. borealis*

Показатель	***A. borealis***
Количество контигов	46 249
Размер сборки (в млн. п. н.)	66,79
N50 по скаффолдам (тыс. п. н.)	649
Место взятия образца	Россия

Выявление повторяющихся последовательностей было проведено при помощи программы RepeatModeler, предназначенной для поиска повторов *de novo*[27]. Для продолжения работы с повторяющимися последовательностями был использован инструмент TEclass classifier [28].

2.2 Программное средство TEclass

Существуют различные категории программ в соответствии с их методологией и характером анализируемых повторов, которые они могут идентифицировать. Выбор метода для поиска повторяющихся последовательностей зависит от уровня знаний о типах повторов, встречающихся в исследуемом геноме.

Можно провести поиск конкретных типов повторов (например, поиск только коротких тандемных повторов, поиск ретротранспозонов), поиск повторов, имеющих характерные структурные особенности, и поиск совершенно новых и неизвестных элементов только на основе их повторяющегося характера [29].

Большое количество секвенированных геномов требовало разработки программного обеспечения, которое восстанавливало консенсусные последовательности транспозонов и других повторяющихся элементов. TEclass

– инструмент, позволяющий классифицировать неизвестные элементы на четыре основных таксономических порядка, которые отражают их режим транспонирования: ДНК-транспозоны, длинные концевые повторы (LTR), длинные диспергированные элементы (LINE) и короткие диспергированные элементы (SINE). TEclass использует механизм поддержки машинного обучения (SVM) для классификации на основе частот олигомеров. Он достигает 90-97% точности в классификации новых ДНК и повторений LTR и 75% для LINE и SINE [30].

Программа анализирует повторы разных размеров: 0-600, 601-1800, 1801-4000 и > 4000 п.н.о. и использует LIBSVM [31] в качестве механизма векторной машины поддержки (SVM) с ядром Gaussian. Процесс классификации является двоичным, со следующими шагами (рис. 6): ориентация forward и reverse, ДНК и ретротранспозон, LTR по сравнению с non-LTR для ретроэлементов, LINEs против SINEs для повторов без LTR. Последний шаг выполняется только для повторов с длиной ниже 1800 п.о., потому что неизвестно о SINE длиной более 1800 пар оснований.

На каждом этапе классификации последовательность TEs представляется в виде вектора частот олигомера, который использовался в качестве входа для SVM [32].

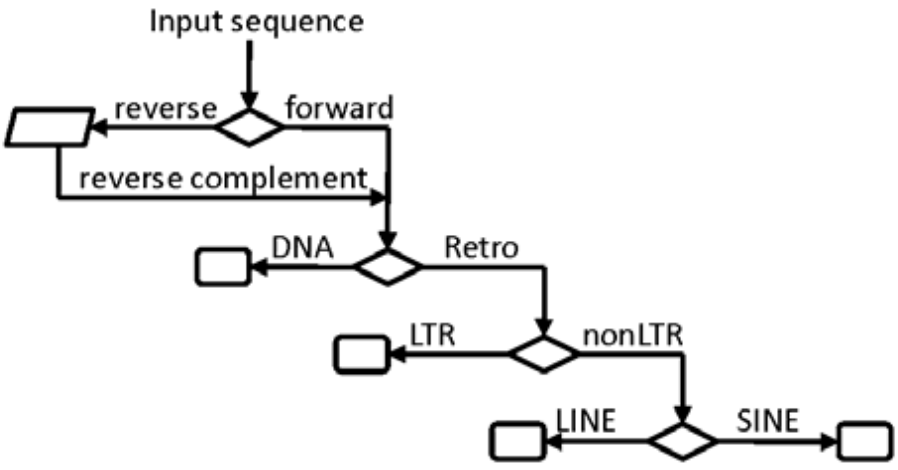

Рисунок 7 — Стадии классификации TEclass

3 РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЯ

3.1 Результаты работы RepeatModeler

Для анализа повторяющихся последовательностей в геномах *P. niemelaei*, *P. chrysoloma* и *Armillaria borealis* были использованы программы, основанные на *de novo* подходах, а также программы, использующие библиотеки повторов (RepeatMasker, RepeatModeler). В таблице А.1 представлены результаты классификации повторов, найденных в геноме *P. niemelaei.*

В геноме *Porodaedalea niemelaei* изначально было найдено 158 повторов, 122 (77, 2%) из них изначально были неклассифицированы. Аналогичная ситуация сложилась и с *Porodaedalea chrysoloma* – из 122 повторов 94 (77,1%) не принадлежали к каким-либо из известных семейств. В таблице А.2 представлены результаты классификации найденных повторов для *Porodaedalea chrysoloma*.

В геноме *Armillaria borealis* было выявлено больше всего повторяющихся последовательностей – 886. Однако, 839 (94,7%) из них изначально остались нераспознанными. В таблице А.3 представлены результаты классификации повторов, найденных в геноме *A. borealis*

Из классифицированных повторов самыми часто встречаемыми оказались LTR-ретротранспозоны, а именно представители суперсемейств *Ty1/Copia* и *Ty3/Gypsy*.

Были выявлены следующие типы повторов: LTR-ретротранспозоны *(Copia, Gypsy, Pao*), ДНК-транспозоны (*hAT-hATw, CMC-EnSpm, TcMar-Pogo,* TcMar-Sagan), LINE-ретротранспозоны (*R1, Tad1*), RC-транспозон *Helitron*, сателлитные повторы.

3.2 Результаты работы TEclass classifier

TEclass classifier позволил распределить неклассифицированные ранее последовательности на 4 основных таксономических порядка. Результаты представлены в таблицах А.4, А.5, А.6.

Результаты сравнения полученных данных по *Porodaedalea niemelaei* и *Porodaedalea chrysoloma* показали значительное сходство как найденных повторов, так и их пропорционального соотношения. Из классифицированных повторов самыми часто встречаемыми оказались DNA-ретротранспозоны, а именно представители семейств TcMar-Sagan, TcMar-Ant1, CMC-EnSpm и LTR-ретротранспозоны (чаще всего транспозоны семейств Copia и Gypsy).

Однако, в сравнении с геномомами *Porodaedalea niemelaei* и *Porodaedalea chrysoloma* у *Armillaria borealis* наблюдаются значительные расхождения.

У рода *Porodaedalea* 40-45 % повторов пришлось на транспозоны II типа (DNA), у *Armillaria* их оказалось всего около 10 %.

Длинные концевые повторы *у Armillaria* составляют 83 %, у *P. nimelaei* и *P. chrysoloma* их 28 % и 34 %, соответственно.

На рисунке 7 предсавлена изменчивость числа повторов между видами *P. niemelaei, P. chrysoloma* и *Armillaria borealis*

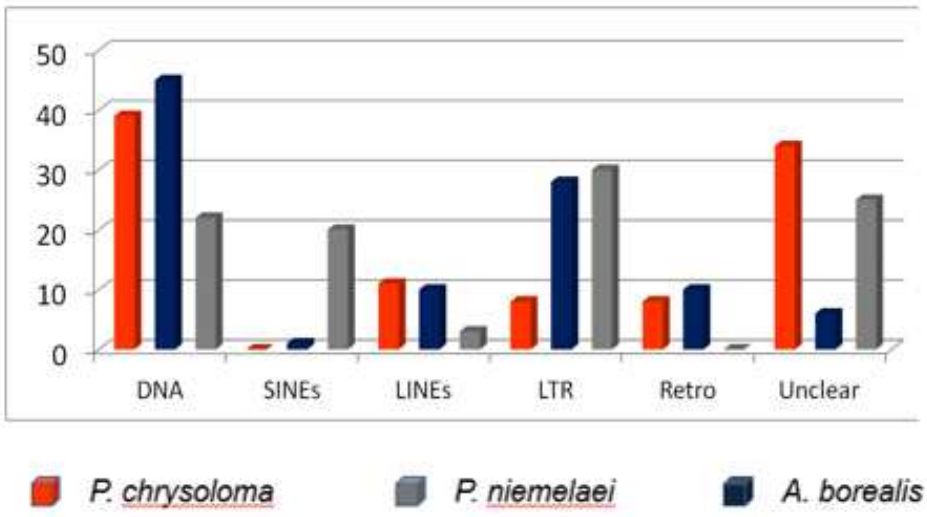

Рисунок 8 — Изменчивость числа повторов между тремя видами

В дальнейшем исследование повторяющихся последовательностей в геномах базидиомицетов и их сравнительный анализ поможет более точно определить таксономическое положение этих видов и выявить эволюционные взаимоотношения.

ВЫВОДЫ

1) Получены библиотеки повторов для трёх видов грибов. *Porodaedalea niemelaei* содержит 158 повторов, *P. chrysoloma* – 122, а *Armillaria borealis* – 886.
2) Проведена классификация ранее неизвестных геномных повторов в *Porodaedalea chrysoloma, Porodaedalea niemelaei* и *Armillaria borealis*;
3) У рода *Porodaedalea* 40-45 % повторов пришлось на транспозоны II типа (DNA), у *Armillaria* их оказалось всего около 10 %.
4) Длинные концевые повторы *у Armillaria* составляют 83 %, у *P. nimelaei* и *P. chrysoloma* их 28 % и 34 %, соответственно.

Список сокращений

1. п.н.о., п.н. -пары нуклеотидных оснований
2. н.о.- нуклеотидные основания
3. ORF- open reading frame (открытые рамки считывания)
4. TEs-Transposable elements (подвижные элементы)
5. Mb – мегабаза (миллион пар оснований)

Список использованных источников

1 Kubiak, M. R.,. Protein-Coding Genes' Retrocopies and Their Functions / M. R Kubiak , I. Makałowska //Viruses. – 2017. – Т. 9. – №. 4. – С. 80.

2 Пармасто Э. Х. Проблема вида у грибов // Проблемы вида и рода у грибов. Таллин. – 1986. – С. 9-29.

3 Sipos, G., Genome expansion and lineage-specific genetic innovations in the forest pathogenic fungi *Armillaria* /G. Sipos //Nature ecology & evolution. – 2017. – Т. 1. – №. 12. – С. 1931.

4 Hirsch, C. D, Transposable element influences on gene expression in plants /C. D. Hirsch, N. M. Springer // *Biochimica et Biophysica Acta* Gene Regulatory Mechanisms. - 2017. - 1860(1).

5 Патрушев, Л. И., Проблема размера геномов эукариот / Л. И. Патрушев, И. Г. Минкевич // Успехи биологической химии. – 2007. – Т. 47. – С. 293–300.

6 Чалей, М. Б., Исследование феномена скрытой периодичности в геномах эукариотических организмов / М. Б. Чалей, В. А. Кутыркин, Е. И. Теплухина // Математическая биология и биоинформатика. – 2013. – Т. 8, № 2 – С. 481.

7 Feschotte, C., Transposable elements and the evolution of regulatory networks /C. Feschotte// Nature Reviews Genetics. – 2008. – Т. 9. – №. 5. – С. 397-405.

8 Wicker, T., A unified classification system for eukaryotic transposable elements / F. Sabot, A. Hua-Van, , J. Bennetzen, //Nature Reviews Genetics. – 2007. – Т. 8. – №. 12. – С. 973.

9 Kapitonov, J., Rolling-circle transposons in eukaryotes /J. Kapitonov// Proc Natl Acad Sci USA – 2001. – Т. 98. – № 15. С. 8714–8719

10 Kubiak, M. R., Protein-Coding Genes' Retrocopies and Their Functions /M. R. Kubiak, I. Makałowska// Viruses. – 2017. – Т. 9. – №. 4. – С. 80.

11 Dennenmoser, S., Copy number increases of transposable elements and protein coding genes in an invasive fish of hybrid origin / S. Dennenmoser, F. J. Sedlazeck, E. Waszkiewicz// Molecular Ecology. – 2017. – Т. 26. – №. 18. – С. 4712-4724.

12 Гогвадзе, Е. В., Влияние ретроэлементов семейств L1 и HERV-K (HML-2) на структуру генома и функционирование близлежащих генов: дис. к.б.н. : 03.00.03 /Е. В. Гогвадзе //Гогвадзе Елена Владимировна – Москва, 2007. – 8 с.

13 Havecker, E. R., The diversity of LTR retrotransposons /E. R. Havecker, X. Gao, D. F. Voytas //Genome biology. – 2004. – Т. 5. – №. 6. – С. 225.

14 Сормачева, И. Д., LTR-ретротранспозоны растений / И. Д. Сормачева, А. Г. Блинов // Вавиловский журнал генетики и селекции. – 2011. – Т. 15. – № 2. – С. 352-360.

15 Eickbush, T. H., The diversity of retrotransposons and the properties of their reverse transcriptases / T. H. Eickbush, V. K. Jamburuthugoda // Virus Res. Author manuscript. – 2008. – Т. 134. – №. 1-2. – С. 221-234.

16 Wicker, T., A unified classification system for eukaryotic transposable elements / F. Sabot, A. Hua-Van, , J. Bennetzen, //Nature Reviews Genetics. – 2007. – Т. 8. – №. 12. – С. 973.

17 Timberlake, W. E., Low repetitive DNA content in *Aspergillus nidulans* /W. E. Timberlake//Science. – 1978. – Т. 202. – №. 4371. – С. 973-975.

18 Kessler, M.M., Systematic discovery of new genes in the Saccharomyces cerevisiae genome /M. M. Kessler//Genome research. – 2003. – Т. 13. – №. 2. – С. 264-271.

19 Dusenbery, R. L., Characterization of the genome of *Phycomyces blakesleeanus* /R. L. Dusenbery//Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis. – 1975. – Т. 378. – №. 3. – С. 363-377.

20 Wöstemeyer, J.,. Structural organization of the genome of the zygomycete *Absidia glauca*: evidence for high repetitive DNA content /J. Wöstemeyer,. Burmester A. //Current genetics. – 1986. – Т. 10. – №. 12. – С. 903-907.

21 Daboussi, M. J.,. Transposable elements in filamentous fungi / M. J. Daboussi, P. Capy //Annual Reviews in Microbiology. – 2003. – Т. 57. – №. 1. – С. 275-299.

22 Lerat, E., Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs /E. Lerat//Heredity. – 2010. – Т. 104. – №. 6. – С. 520.

23 Jurka, J., Repbase Update, a database of eukaryotic repetitive elements /J. Jurka, V. V. Kapitonov, A.Pavlicek, P.Klonowski, Kohany, O., & Walichiewicz,//Cytogenetic and genome research. – 2005. – Т. 110. – №. 1-4. – С. 462-467.

24 Jurka, J., CENSOR—a program for identification and elimination of repetitive elements from DNA sequences /J. Jurka, P. Klonowski, V. Dagman,Pelton//Computers & chemistry. – 1996. – Т. 20. – №. 1. – С. 119-121.

25 Pevzner, P. A., De novo repeat classification and fragment assembly / H. Tang,G. Tesler //Genome research. – 2004. – Т. 14. – №. 9. – С. 1786-1796.

26 Пятков, М. И., Разработка спектрального подхода к поиску протяженных повторяющихся последовательностей в геномах: дис. к.ф.-м.н.: 03.01.02 / Пятков Максим Иванович. – Пущино, 2013. – С. 14

27 Smit ,A., RepeatModeler-1.0. 5 // A. Smit, R. Hubley /Institute for Systems Biology. – 2012.

28 Abrusán, G., TEclass—a tool for automated classification of unknown eukaryotic transposable elements /G. Abrusán //Bioinformatics. – 2009. – Т. 25. – №. 10. – С. 1329-1330.

29 Lerat, E., Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs /E. Lerat //Heredity. – 2010. – Т. 104. – №. 6. – С. 520.

30 Andrieu, O., Detection of transposable elements by their compositional bias /O. Andrieu //BMC bioinformatics. – 2004. – Т. 5. – №. 1. – С. 94.

31 Chang, C. C., LIBSVM: a library for support vector machines /C. C. Chang, Lin C. J. //ACM transactions on intelligent systems and technology (TIST). – 2011. – Т. 2. – №. 3. – С. 27.

32. Lerat, E., Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs /E. Lerat //Heredity. – 2010. – Т. 104. – №. 6. – С. 520.

33. Xiong, Y., Origin and evolution of retroelements based upon their reverse transcriptase sequences / Y. Xiong, T. H. Eickbush //The EMBO journal. – 1990. – Т. 9. – №. 10. – С. 3353.

34. Horns, F.,Patterns of repeat-induced point mutation in transposable elements of *Basidiomycete* fungi /F. Horns//Genome biology and evolution. – 2012. – Т. 4. – №. 3. – С. 240-247.

35. Maumus, F., Deep investigation of Arabidopsis thaliana junk DNA reveals a continuum between repetitive elements and genomic dark matter/ F. Maumus, H. Quesneville //PLoS One. – 2014. – Т. 9. – №. 4. – С. e94101.

36. Dhillon, B.,The landscape of transposable elements in the finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* /B. Dhillon//BMC genomics. – 2014. – Т. 15. – №. 1. – С. 1132.

37. Karaoglu, H., Survey of simple sequence repeats in completed fungal genomes /H. Karaoglu//Molecular Biology and Evolution. – 2005. – Т. 22. – №. 3. – С. 639-649.

ПРИЛОЖЕНИЕ

Таблица А.1 – Классификация найденных повторов для *P. niemelaei*

Тип повтора	Количество
CMC-EnSpm	3
Copia	5
Gypsy	12
hAT-hATw	2
Helitron	1
Pao	1
R1	1
RTE-X	1
SSR	2
Tad1	1
TcMar-Ant1	1
TcMar-Fot1	1
TcMar-Pogo	1
TcMar-Sagan	2
Ngaro	—
Simple	2
Classified	**36 (22,8 %)**
Unclassified	**122 (77,2 %)**
Total	**158**

Таблица А.2– Классификация найденных повторов для *P. chrysoloma*

Тип повтора	Количество
CMC-EnSpm	1
Copia	5
Gypsy	13
hAT-hATw	2
Helitron	—
Pao	—
R1	—
RTE-X	—
SSR	2
Tad1	1
TcMar-Ant1	2
TcMar-Fot1	—
TcMar-Pogo	1
TcMar-Sagan	1
Ngaro	—
Simple	—
Classified	**28 (22,9%)**
Unclassified	**94 (77,1%)**
Total	**122**

Таблица А.3 – Классификация найденных повторов для *A. borealis*

Тип повтора	Количество
CMC-EnSpm	—
Copia	5
Gypsy	32
hAT-hATw	—
Helitron	—
Pao	1
R1	—
RTE-X	—
SSR	—
Tad1	5
TcMar-Ant1	—
TcMar-Fot1	—
TcMar-Pogo	—
TcMar-Sagan	—
Ngaro	3
Simple	1
Classified	**47 (5,3%)**
Unclassified	**839 (94,7%)**
Total	**886**

Таблица А.4 – Классификация неизвестных повторов для *P. niemelaei*

Тип повтора	**Количество**	**Количество в %**
DNA	47	39
LTR	41	34
LINE	14	11
Retrotransposons	10	8
Non-LTR	—	0
Unclear	10	8
Total	122	100

Таблица А.5 – Классификация неизвестных повторов для *P. niemelaei*

Тип повтора	**Количество**	**Количество в %**
DNA	71	45
LTR	45	28
LINE	16	10
Retrotransposons	9	6
Non-LTR	1	1
Unclear	16	10
Total	158	100

Таблица А.6– Классификация неизвестных повторов для *A. borealis*

Тип повтора	**Количество**	**Количество в %**
DNA	79	9
LTR	735	83
LINE	27	3
Retrotransposons	27	3
Non-LTR	0	0
Unclear	18	2
Total	886	100

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ
Заведующий кафедрой
_____ В. А. Кратасюк
подпись инициалы, фамилия
«9» июня 2018г.

БАКАЛАВРСКАЯ РАБОТА

06.03.01 Биология
06.03.01.07 Биофизика

Анализ повторяющихся последовательностей в геномах *Porodaedalea niemelaei*, *P. chrysoloma* и *Armillaria borealis*

Научный руководитель _____ 14.06.2018 г. Ю. А. Путинцева
подпись, дата инициалы, фамилия

Научный консультант _____ 14.06.2018 д.ф.-м.н., в.н.с. М. Г. Садовский
подпись, дата должность, ученая степень инициалы, фамилия

Выпускник _____ 14.06.2018 А. И. Аксёнова
подпись, дата инициалы, фамилия

Красноярск 2018