На правах рукописи

Киров Илья Владимирович

ОСОБЕННОСТИ ОРГАНИЗАЦИИ ПОВТОРЯЮЩИХСЯ ЭЛЕМЕНТОВ ГЕНОМОВ РАСТЕНИЙ, ВЫЯВЛЕННЫЕ С ПОМОЩЬЮ НОВЫХ ОМИКСНЫХ ПОДХОДОВ

1.5.7 – Генетика

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
доктора биологических наук

Москва - 2024

Работа выполнена в Федеральном государственном бюджетном научном учреждении «Всероссийский научно-исследовательский институт сельскохозяйственной биотехнологии» (ФГБНУ ВНИИСБ), г. Москва

Научный консультант:
СОЛОВЬЕВ Александр Александрович
доктор биологических наук, профессор, профессор РАН, заместитель директора Федерального государственного бюджетного учреждения «Всероссийский центр карантина растений», Московская область, г.о Раменский, р.п. Быково

Официальные оппоненты:
КУЛУЕВ Булат Разяпович
доктор биологических наук, заведующий лабораторией геномики растений Института биохимии и генетики – обособленного структурного подразделения Федерального государственного бюджетного научного учреждения Уфимского федерального исследовательского центра Российской академии наук, г. Уфа

САЛИНА Елена Артемовна
доктор биологических наук, профессор, заведующий отделом молекулярной генетики растений Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», г. Новосибирск

САМСОНОВА Мария Георгиевна
доктор биологических наук, заведующий Научно-исследовательской лабораторией математической биологии и биоинформатики Федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого», г. Санкт-Петербург

Ведущая организация:
Федеральное государственное бюджетное учреждение науки Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук, г. Новосибирск

Защита диссертации состоится «__» _____ 2024 года в ____ часов на заседании диссертационного совета 24.1.088.01 в Федеральном государственном бюджетном учреждении науки Институт общей генетики им. Н. И. Вавилова Российской академии наук по адресу: 119991, Москва, ул. Губкина, 3.
С диссертацией и авторефератом можно ознакомиться в библиотеке и на сайте Института www.vigg.ru, тел. 8-499-135-14-31, e-mail: dissovet@vigg.ru

Автореферат диссертации разослан «_____»_____2024 года.

Учёный секретарь
диссертационного Совета,
доктор биологических наук Горячева И. И.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы и степень разработанности проблемы. Значительная часть геномов растений представлена различными видами повторяющихся элементов, включая высококопийные тандемно-организованные повторы и мобильные элементы. Долгое время повторяющиеся элементы генома рассматривали как функционально незначимую и «мусорную» ДНК. Но с появлением новых способов секвенирования и молекулярно-биотехнологических методов, получено всё больше доказательств ключевой роли повторяющихся элементов в широком спектре биологических и эволюционных процессов. Так, например, тандемно-организованные повторы генома растений являются главными структурными элементами центромер, теломер и гетерохроматина, и принимают непосредственное участие в функции этих важных морфологических структур хромосом. Данные последних исследований по сравнению геномов сотен сортов разных культурных растений свидетельствуют, что активность мобильных элементов способствовала появлению многих ценных признаков современных сельскохозяйственных растений. Новые инсерции мобильных элементов вносят огромный вклад в генетическую вариабельность, в появление новых генов и транскрипционных программ и играют важнейшую роль в адаптации растений к новым экологическим нишам и меняющимся условиям среды. Но, несмотря на это, повторяющиеся элементы остаются самой малоизученной частью генома растений. Новые технологии, такие как секвенирование длинными ридами PacBio и Oxford Nanopore Technology (ONT), позволяют решить проблемы сборки повторяющихся элементов, что долгое время ограничивало их изучение. Поэтому создание новых биоинформатических и молекулярных подходов для системного изучения повторяющихся элементов с использованием современных омиксных данных (эпигеномика, транскриптомика, протеомика, циркуломика) и методов секвенирования (секвенирование длинными ридами) – это актуальная задача современной биологии, на решение которой и направлена данная работа.

Кроме фундаментальных вопросов, последовательности повторяющихся элементов представляют большой практический интерес. Прежде всего он связан с использованием знаний о повторах генома для изучения вариабельности геномов, возникающей в естественных условиях и в условиях биотехнологического размножения растений и селекции. К этой области относятся такие актуальные направления современной биотехнологии и генетики как: 1) создание цитогенетических маркеров на основе тандемных повтыоров и изучение хромосомной вариабельности; 2) дизайн молекулярных маркеров, вовлекающих полиморфизм повторяющихся элементов; 3) поиск и аннотация повторяющихся элементов, как важной части современных проектов по секвенированию и сборке геномов растений, что позволяет ускорять процесс аннотации генов и проводить интегрирование собранных сиквенсов с физическими хромосомами; 4) инсерционный мутагенез с использованием мобильных элементов.

Мобильные элементы, как показывают последние исследования, являются драйверами эволюции и селекции. В связи с этим актуальным практическим направлением современной биотехнологии и генетики является создание методов

для контролируемой активации нативных мобильных элементов генома растений. Успехи в этой области позволят расширить генетическое разнообразие, а также ускорить селекцию и развитие функциональной геномики. Для создания таких методов необходимы глубокие знания как о системах сайленсинга мобильных элементов, так и о биологии мобилома и активных мобильных элементах в геномах разных видов культурных растений. За последние двадцать лет были детально изучены системы сайленсинга мобильных элементов растений. Наряду с этим, биология самих мобильных элементов, включая жизненный цикл, взаимодействие с клеточными структурами и процессинг РНК, практически не изучены у растений. Кроме этого, хотя миллионы мобильных элементов были аннотированы в собранных геномах растений, лишь для нескольких десятков была доказана транспозиционная активность. Крайний недостаток знаний об активных мобильных элементах растений существенно мешает прогрессу в области биологии мобилома и контролируемой активации мобильных элементов. Поэтому разработка и внедрение новых методов полногеномного анализа активности мобилома и идентификации новых инсерций мобильных элементов являются критически необходимыми для современной биологии растений. Это является одной из задач, на решение которой направлено это диссертационное исследование.

Цель и задачи. Целью данной работы являлось системное изучение повторяющихся элементов (сателлитных повторов и мобильных элементов) геномов разных видов растений (однодольные (*Allium cepa, A. fistulosum, x Triticosecale, Triticum aestivum*), двудольные (*Rosa wichurana, R. gallica, R. rugosa, R. foetida, R. chinensis, Helianthus annuus, Arabidopsis thaliana, Brassica napus*) и мох (*Physcomitrium patens*)) и разработка для этого новых молекулярных и биоинформатических подходов.

Для достижения поставленной цели были решены следующие задачи:

1. Идентификация и цитогеномный анализ новых сателлитных повторов и их использование для разработки хромосомных маркеров разных видов растений.
2. Разработка новых биоинформатических подходов для идентификации и цитогенетического анализа сателлитных повторов геномов растений, используя данные полногеномного нанопорового секвенирования и собранные геномы.
3. Разработка новых молекулярных и биоинформатических подходов для поиска и характеристики инсерций мобильных элементов растений по данным нанопорового секвенирования.
4. Поиск новых активных мобильных элементов и изучение их геномных и транскриптомных особенностей у однодольных и двудольных видов растений.
5. Транскриптомный анализ сателлитных повторов и мобильных элементов растений.
6. Изучение структуры и состава внехромосомных кольцевых молекул ДНК LTR ретротранспозонов растений, используя нанопоровое секвенирование и полногеномный анализ.
7. Протеомный анализ белков мобильных элементов на примере *Arabidopsis thaliana*.

Научная новизна и практическая значимость работы. В рамках диссертационной работы разработан комплекс биоинформатических и молекулярно-биотехнологических методов, направленных на идентификацию новых повторяющихся элементов геномов растений и изучение генетической вариабельности, обусловленной этими элементами, включая поиск новых инсерций мобильных элементов. Использование этих методов позволило впервые изучить на геномном и постгеномном уровнях повторяющиеся элементы генома как важных сельскохозяйственных (подсолнечник, гречиха, лук, тритикале, роза), так и модельных (арабидопсис, мох (*Physcomitrium patens*)) растений. Биоинформатические программы (NanoCasTE и nanotei) и молекулярный метод CANS, разработанные для детекции инсерций мобильных элементов в геноме, позволили впервые установить закономерности распределения соматических инсерций ретротранспозонов в геноме. Полученные сведения впервые позволили показать связь между частотой инсерций, определёнными хромосомными (центромерные регионы) регионами, и эпигенетическими и транскриптомными особенностями генома. Разработанные методы детекции новых инсерций мобильных элементов могут быть использованы также в биотехнологии растений для детекции Т-ДНК и ускорения процесса создания новых генотипов культурных растений, несущих новые функционально и фенотипически значимые инсерции.

В работе впервые выделен и изучен набор новых высококопийных тандемных повторов для разных видов растений (*Allium fistulosum*: AfiCen1K; *Allium cepa*: TR2CL37, *Rosa wichurana* CL8, CL24; *Rosa chinensis*: CL226; 19 повторов для *Physcomitrium patens*), что представляет уникальные инструменты для молекулярных и эволюционных исследований генома растений. Важным результатом данной работы стала разработка новой программы, DRAWID, для ускорения цитологического и молекулярно-цитогенетического анализа, и показано, что данная программа облегчает базовый анализ кариотипа для прикладных цитологических исследований, а также анализ кариотипа после гибридизации *in situ*, что позволяет эффективнее изучать хромосомную организацию повторяющихся последовательностей генома.

В работе были идентифицированы и детально изучены на митотических и мейотических хромосомах повторяющиеся последовательности и организация центромер *A. cepa* и *A. fistulosum*. Полученные результаты впервые показывают, что центромеры хромосом *A. fistulosum* содержат длинный (~1,25 т.п.н.) тандемный повтор AfiCen1K, а также вставки ретротранспозонов и ДНК органелл. Полученные сведения о сателлитных повторах позволили впервые провести дизайн системы хромосомных маркеров для видов луковых и роз, что может быть использовано для интегрирования физических и генетических карт, а также улучшения полногеномной сборки. Полученные сведения позволили провести аннотацию центромерных последовательностей в собранном геноме *Rosa chinensis* и провести интегрирование хромосомных и геномных карт в рамках международного проекта по секвенированию генома.

В рамках транскриптомного анализа повторяющихся элементов впервые показано, что десятки LTR ретротранспозонов и сателлитные повторы

экспрессируются как в нормальных, так и в стрессовых условиях. Более того, было показано, что экспрессия сателлитных повторов характерна для филогенетически удалённых растений, включая лук (*Allium cepa)* и мох (*Physcomitrium patens)*. В работе показаны отличительные особенности геномной и транскриптомной организации экспрессирующихся ретротранспозонов, включая более раннее время инсерции, специфический филогенетический состав, преимущественное кодирование транскриптами GAG белка и слабая связь экспрессии с мобильной активностью. Полученные сведения вносят важный вклад в изучение фундаментальной проблемы роли мобильных элементов в формировании транскриптома растений.

Важнейшим результатом данной работы является адаптирование нанопорового секвенирования для полногеномного анализа внехромосомных кольцевых ДНК (вкДНК) растений. Используя данный подход для разных видов растений (рапс и арабидопсис), мы впервые изучили состав и структуру вкДНК и показали, что вкДНК LTR ретротранспозонов представляет пул гетерогенных по структуре молекул. Благодаря новому подходу, список известных мобильных элементов с доказанной мобильной активностью был существенно расширен новыми элементами генома тритикале (ретротранспозон 'MIG'), подсолнечника (ретротранспозоны 'Gagarin' и 'SUNTY3'), рапса (семейство ретротранспозонов 'Antares') и арабидопсиса ('TR-GAG' элемент). Новые открытые элементы позволят изучить биологию мобилома не только у модельных растений, но и у культурных растений с большими и сложными геномами. Кроме этого, полученные результаты создают научную основу для дальнейшего использования методов контролируемой активации мобилома для создания инсерционных коллекций, а также изучения влияния новых инсерций на структуру генома и эпигенома растений.

Используя современный подход, комбинирующий прямое секвенирование РНК и масс-спектрометрический анализ протеома на уникальном растительном материале с активными мобильными элементами – линией с мутацией в гене DDM1 (*ddm1*), впервые показано, что мобильные элементы могут кодировать белки разной природы, включая белки с известной функцией (например, транспозаза и белок *Env*), а также неизвестные белки, функции которых только предстоит понять в будущем.

Основные положения, выносимые на защиту:

1. Идентифицированы новые сателлитные повторы *Allium fistulosum* (AfiCen1K), *Allium cepa* (TR2CL37), *Rosa wichurana* (CL8, CL24), *Rosa chinensis* (CL226) и 19 повторов для *Physcomitrium patens*, которые составляют гетерохроматин или центромерные регионы хромосом и подходят для цитогенетического маркирования хромосом изучаемых видов.
2. Экспрессия сателлитных повторов с образованием преимущественно некодирующих РНК наблюдается у филогенетически отдаленных видов

растений, включая покрытосеменные растения (*Allium cepa*, *A. fistulosum*) и мох (*Physcomitrium patens*).

3. Разработанные новые биоинформатические программы pyTanFinder и nanoTRF позволяют идентифицировать в геноме и проводить анализ новых сателлитных повторов на основе данных собранных сиквенсов геномов и ридов от нанопорового секвенирования ДНК, соответственно.
4. Разработанная новая программа, DRAWID, позволяет проводить базовые измерения хромосом и определять положение сигналов на них и строить идиограммы разных типов, облегчая анализ кариотипа для прикладных цитологических и молекулярно-цитогенетических (гибридизация *in situ*) исследований, что позволяет эффективнее изучать хромосомную организацию повторяющихся последовательностей генома.
5. Десятки LTR ретротранспозонов однодольных и двудольных растений экспрессируются в нормальных и стрессовых условиях, формируя ретротранскриптом, в котором значительная доля транскриптов кодируют GAG белки мобильных элементов.
6. Распределение новых инсерций LTR ретротранспозонов *ONSEN* и *EVD* в геноме *Arabidopsis thaliana* неслучайно и связано с эпигенетическими и транскрипционными характеристиками генома, как показано с помощью разработанных новых молекулярных (CANS) и биоинформатических подходов (NanoCasTE, nanotei).
7. Пул гетерогенных по структуре и композиции внехромосомных кольцевых ДНК LTR ретротранспозонов растений образуется под действием эпигенетического стресса (воздействие альфа-аманитина и зебуларина) в комбинации с тепловым стрессом.
8. Идентифицированы новые активные мобильные элементы генома тритикале (ретротранспозон 'MIG'), подсолнечника (ретротранспозоны 'Gagarin' и 'SUNTY3'), рапса (семейство ретротранспозонов 'Antares') и арабидопсиса ('TR-GAG' элемент), которые могут быть использованы для изучения закономерностей формирования мобилома как у модельных, так у культурных растений с большим и сложным геномом.
9. Мобильные элементы кодируют не только канонические белки, но и белки с неизвестной функцией.

Личный вклад автора в проведение исследования. Диссертант принимал непосредственное участие в формировании концепции диссертационной работы, формулировке целей и задач, а также в планировании и проведении экспериментов. Основные положения и выводы диссертационной работы сформулированы автором. Экспериментальные данные, представленные в диссертации, получены лично диссертантом или в соавторстве с сотрудниками, работавшими совместно с автором в процессе выполнения исследований. Биоинформатические программы, описанные в диссертации, были разработаны непосредственно автором.

Апробация работы. Материалы работы были представлены в виде стендовых и устных докладах на российских и международных научных конференциях, среди которых: XIX Всероссийский симпозиум «Структура и функции клеточного ядра»

(г. Санкт-Петербург, 2024), VII Международная научная конференция «Генетика, Геномика, Биоинформатика и Биотехнология растений» (PlantGen 2023, Казань) и Школа молодых учёных (PlantGenSchool 2023, Казань), I и II Международная молодежная конференция "Генетические и радиационные технологии в сельском хозяйстве" (2022 и 2023 гг., Обнинск), Всероссийская школа-конференция «Клеточные и геномные технологии для совершенствования сельскохозяйственных животных» (Санкт-Петербург-Пушкин, 2022 и 2023 гг.), OpenBio-2022 (онлайн, 2022), XXII-я Международная научная конференция молодых ученых «Биотехнология в растениеводстве, животноводстве и сельскохозяйственной микробиологии», посвященная академику РАСХН Г.С. Муромцеву, 5th Uppsala Transposon Symposium (онлайн, 2021), CSHL Virtual Transposable Elements Meeting (онлайн, 2020), 12th European Cytogenomics Conference (Salzburg, Austria, 2019 г.), Plant Genome Evolution Conference (Испания, 2019 г.).

Публикации. Автором опубликовано 42 научных статьи в журналах, рекомендованных ВАК, из них по теме диссертации: 18 научных статей (11 статей Q1 WoS) и 1 патент на изобретение.

Структура и объём диссертационной работы. Диссертация изложена на 266 страницах, содержит 79 рисунков, 14 таблиц и состоит из введения, обзора литературы, описания материалов и методов, результатов и обсуждения, заключения и списка цитированной литературы, содержащего 475 источника.

СОДЕРЖАНИЕ РАБОТЫ

2. ОБЗОР ЛИТЕРАТУРЫ

Представлен обзор литературы по классификации повторяющихся элементов генома, включая сателлитные повторы и мобильные элементы генома (2.1). Описаны особенности геномной организации, структуры и жизненного цикла LTR ретротранспозонов растений (2.2), а также приведена их современная классификация. Подробно рассмотрена тема ретротранскриптома растений (2.3), закономерности организации, состав и методы изучения. Проанализированы имеющиеся данные по мобилому растений (2.4). В обзоре также дан детальный современных методов идентификации сателлитных повторов в геномах растений (2.5).

3. МАТЕРИАЛЫ И МЕТОДЫ

Работа выполнена с использованием методов цитогенетического, молекулярно-биологического, протеомного, транскриптомного, статистического и биоинформацтического анализа. Описан использованный растительный материал (3.1), методики выделения геномной ДНК (3.2), внехромосомных кольцевых ДНК (3.17). Изложены использованные биоинформатические подходы для характеристики репитома (3.3) и идентификации тандемных повторов с помощью коротких ридов, а также описаны RNAseq анализ (3.4), идентификация экспрессирующихся ретротранспозонов подсолнечника (3.13), оценка времени инсерции (3.14), расчёт расстояния между генами и RTE (3.15), анализ мобилома (3.16, 3.21), предсказание длинных некодирующих РНК (3.18) и масс-спектрометрический анализ (3.19). Приведены методики ПЦР (3.5) и комплекса методов для FISH и C-окрашивания (3.6, 3.7, 3.8, 3.9). Даны описания ОТ-ПЦР (3.10), ПЦР-валидации инсерций (3.20), методов приготовления библиотек и нанопоровое секвенирование ДНК (3.11) и РНК (3.12, 3.22).

4. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

4.1 Цитогеномный анализ новых сателлитных повторов растений

4.1.1 pyTanFinder: инструмент для поиска высококопийных тандемных повторов в секвенированных геномах

На сегодня сиквенсы генома доступны для большого числа разных видов живых организмов. Эти данные представляют ценную информацию и для поиска и изучения геномной организации разных повторяющихся элементов, включая мобильные элементы и тандемные (сателлитные) повторы (ТП). Однако тандемные повторы очень часто недостаточно представлены в собранных геномах или входят в состав разных контигов. Это препятствует идентификации высококопийных тандемных повторов, которые представляют большой интерес для изучения и создания молекулярно-цитогенетических маркеров. Для решения этой проблемы была разработана программа pyTanFinder (https://github.com/Kirovez/pyTanFinder). Главная задача pyTanFinder заключается в поиске и оценки копийности тандемных повторов в геномных сборках.

Результатом работы этой программы является FASTA-файл всех тандемных повторов и таблица, содержащая уникальные последовательности ТП с предполагаемой распространенностью в геноме. Кроме того, pyTanFinder также формирует html-отчет, содержащий гистограммы распределения размера мономера ТП и количества связей каждого мономера в отдельном кластере. pyTanFinder был применён для поиска ТП в геномах разных видов живых организмов.

4.1.2 nanoTRF: программа для поиска высококопийных тандемных повторов в «сырых» данных нанопорового секвенирования

Разработка секвенирования Oxford Nanopore (ONT) произвело революцию в области геномики, обеспечив высокопроизводительное секвенирование с длинными прочтениями (ридами) по низкой цене с использованием портативного секвенатора (MinION). В контексте исследования тандемных повторов «сырые» риды ONT обеспечивают основу для секвенирования длинных массивов тандемных повторов и изучения геномной организации. Кроме того, данные ONT также можно использовать для расшифровки эпигенетического профиля тандемных повторов. Было предложено несколько алгоритмов для идентификации ТП в отдельных ридах ONT, такие как TideHunter и NCRF. Однако *de novo* идентификация новых высококопийных ТП из «сырых» данных ONT после секвенирования с низким покрытием генома и их классификация по семействам, а также оценка представленности в геноме все еще затруднены. Тем не менее, этот тип данных быстро накапливается в базах данных. Для поиска ТП в «сырых» данных ONT была разработана программа nanoTRF (https://github.com/Kirovez/nanoTRF). nanoTRF – это алгоритм для *de novo* идентификации, количественной оценки и сборки высококопийных ТП. NanoTRF идентифицирует последовательности ТП и рассчитывает их представленность в геноме на основе входных данных.

Ключевой целью программы является идентификация высококопийных тандемных повторов (ТП) и реконструкция их консенсусных последовательностей. Единственный входной файл для работы программы nanoTRF — это геномные риды Nanopore в форматах fastq или fasta. nanoTRF написан на Python 3 и работает из командной строки. Чтобы проверить работу nanoTRF, было проведено секвенирование Nanopore генома *Deschampsia antarctica*, вида с хорошо охарактеризованным составом сателлитома. Всего было получено 1 452 313 прочтений с N50 1305 п.н. и общим количеством оснований ~4 Gb, что составляет примерно 0,8-кратное покрытие генома *D. antarctica*. Всего с помощью nanoTRF было идентифицировано 27 последовательностей ТП со средней длиной мономера 371 п.н. (19–2022 п.н.).

Затем было проведено сравнение результатов nanoTRF с результатами программным обеспечением TAREAN, которое использует риды Illumina. Для этого были идентифицированы ТП с помощью TAREAN, используя чтения Illumina для *D. antarctica*. Сравнение ТП, обнаруженных с помощью nanoTRF и TAREAN, показало, что 81% (22 из 27 ТП) ТП, обнаруженных с помощью nanoTRF, также были обнаружены с помощью TAREAN. Кроме того, nanoTRF обнаружил 6 ТП, которые не были идентифицированы TAREAN. Основываясь на

доле ТП в геноме по данным TAREAN, 2,44% генома D. *antarctica* занято сателлитной ДНК. Значение хорошо согласуется с результатами nanoTRF (2,42%).

Далее сравнили долю в геноме и длину мономера, рассчитанную для ТП с помощью nanoTRF и TAREAN. Результаты показали хорошую корреляцию (коэффициент корреляции> 0,79, значение p <1,1x10^{-5}) двух подходов (Рисунок 1).

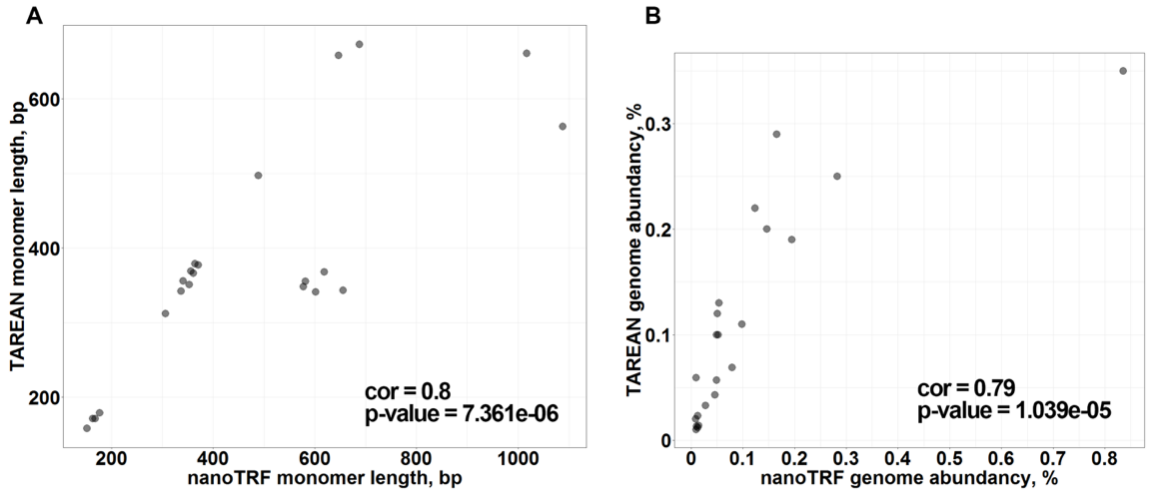

Рисунок 1. Сравнение длины мономера и доли в геноме для ТП идентифицированных nanoTRF и TAREAN.

Выравнивание последовательностей мономеров, собранных с помощью nanoTRF и TAREAN, показало, что большинство последовательностей имеют > 95% идентичности. Кроме того, nanoTRF также обнаружил 5 ТП, которые не были идентифицированы TAREAN, включая clust84 (18 п.н., 0,02%), clust44 (662 п.н., 0,01%), clust64 (123 п.н., 0,02%), clust41 (1964 п.н., 0,01%), clust71 (2022 п.н. , 0,025%).

Таким образом, nanoTRF является новым удобным инструментом для *de novo* идентификации и сборки последовательностей высококопийных тандемных повторов, используя «сырые» данные нанопорового секвенирования.

4.1.3 DRAWID: java программа для измерения параметров хромосом и построение идиограмм хромосом

Одним из инструментов в изучении повторяющихся последовательностей генома на уровне хромосом являются методы молекулярной цитогенетики, включая хромосомный анализ и флуоресцентную гибридизацию *in situ* (FISH). Число, морфология и организация хромосом являются важными параметрами для сравнительных цитогенетических и филогенетических исследований. Однако ни одна из этих программ не может одновременно измерять параметры хромосом и положение FISH и других сигналов на хромосомах.

Для решения этой задачи была разработана программа DRAWID (DRAWing IDiogram). DRAWID представляет собой удобную и бесплатную программу на основе Java, которая облегчает базовый анализ кариотипа, а также анализ кариотипа на основе FISH. DRAWID оснащен интуитивно понятным графическим

интерфейсом. Входными файлами для DRAWID являются файлы изображений (JPEG и PNG) или таблицы данных, созданные самим DRAWID. Выходными данными программы являются таблицы Microsoft Excel (2010), содержащие детали измерений (индекс центромер, соотношение плеч, относительная и абсолютная длина хромосомы и хромосомных плеч, положения и размер сигнала и полосы (если доступны), а также изображения идиограмм, построенные DRAWID. Параметры идиограммы легко настраиваются для подготовки качественного изображения, пригодного для публикации. Кроме того, для облегчения высокопроизводительного кариотипирования программа позволяет собирать данные из разных метафаз и строить среднюю идиограмму с дополнительными элементами, представляющими стандартное отклонение для длин хромосомы и положение центромер. DRAWID v0.26 вместе с руководством по его использованию доступна на GitHub: https://github.com/Kirovez/DRAWID.

Для оценки DRAWID по кариотипированию отдельных метафаз (Рисунок 2A, B), а также набора метафаз были использованы ранее опубликованные данные кариотипирования *Cannabis sativa* Linnaeus, 1753, *Rosa wichurana* Crépin, 1888, *Allium cepa* Linnaeus, 1753 и *A. fistulosum*. Все виды диплоидны, 2n = 2x = 20 (*Cannabis sativa*), 2n = 2x = 14 (*Rosa wichurana*), 2n = 2x = 16 *(Allium cepa* и *A. fistulosum*). Идиограммы для этих видов, построенные с помощью DRAWID, представлены на Рисунке 2.

Таким образом, была разработана новая кроссплатформенная java программа, DRAWID, оснащенная интуитивно понятным графическим интерфейсом, который облегчает базовый анализ кариотипа, а также анализ кариотипа после гибридизации *in situ*, что позволяет эффективнее изучать хромосомную организацию повторяющихся последовательностей генома.

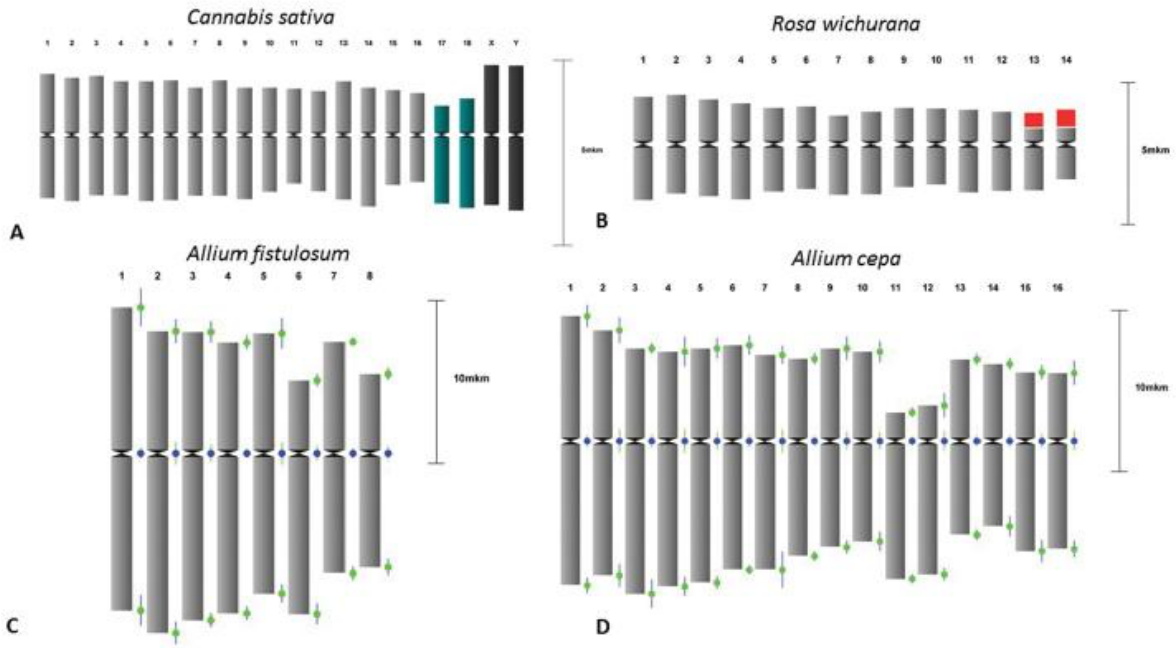

Рисунок 2. Примеры основных измерений кариотипа и построения идиограммы с помощью DRAWID. (A) Хромосомы *C. sativa* (2n = 20); выделены половые (черный цвет) и NOR (зеленый цвет) хромосомы (B) Идиограмма *Rosa wichurana*;

сателлиты на хромосомах 13 и 14 окрашены в красный цвет (C) Идиограмма *Allium fistulosum* после измерения хромосом и применения функции «reduce karyo», объединяющей гомологичные хромосомы для получения 1n идиограммы, показаны столбцы стандартного отклонения (D) Построенная идиограмма *Allium cepa* после измерений 3 метафаз и применения функции «получить среднее карио» для получения средней идиограммы показаны столбцы стандартного отклонения.

4.1.4 Тандемные повторы *A. fistulosum*, ассоциированые с центромерными и гетерохроматиновыми регионами хромосом

Виды *Allium* являются важными овощными культурами, распространенными по всему миру. Однако из-за большого размера генома (8-32 Gb/1C) и высокого содержания повторяющихся последовательностей, геном луковых слабо изучен. Цель этого исследования состояла в анализе репитома и идентификации новых тандемных повторов генома *A. fistulosum*, которые можно использовать в качестве цитогенетических маркеров.

Для характеристики репитома *A. fistulosum* была использована программа RepeatExplorer и геномные риды Illumina *A. fistulosum* из NCBI (SRX268217). Для идентификации кластеров, соответствующих тандемным повторам, был проведён поиск кластеров глобулоподобной (CL2, CL5, CL6, CL7 и CL58) или кольцеобразной (CL36) формы, которые типичны для тандемно-организованных повторяющихся последовательностей. Контиги из CL2, CL5, CL6 и CL7 показали сходство с известными субтеломерными повторами *A. fistulosum*, а также повторами HAT58 и CAT36, которые не были детально изучены ранее. FISH с HAT58 на митотических хромосомах *A. fistulosum* выявила, что локус HAT58, расположенный на хромосоме 7, является полиморфным между растениями, и были идентифицированы три типа FISH-паттернов на хромосоме 7: (1) отсутствие сигналов на обоих гомологах хромосомы 7; (2) наличие сигнала на одном гомологе хромосомы 7; 3) наличие сигналов на обоих гомологах хромосомы 7. Совместное использование FISH и C-бэндинга и последующий статистический анализ измерений с помощью DRAWID подтвердили совместную локализацию флуоресцентных сигналов HAT58 и соответствующих C-бэндов на плечах хромосом 6, 7 и 8. Результаты показывают, что HAT58 является составной частью интеркалярного гетерохроматина хромосом 6, 7 и 8 *A. fistulosum*.

FISH с зондами CAT36 с хромосомами *A. fistulosum* выявил сигналы в прицентромерных областях хромосомы 5 (RL 11 ± 1,5; CI 47,9 ± 1,9) и хромосомы 6. Мы обнаружили, что CAT36 находится в прицентромерной области хромосом 5 и 6 *A. fistulosum*. Двухцветная FISH с Afi11 и CAT36 показала, что сигналы от Afi11 (красный) и CAT36 (зеленый) перекрываются на хромосомах 5 и 6 (Рисунок 3A; вставка).

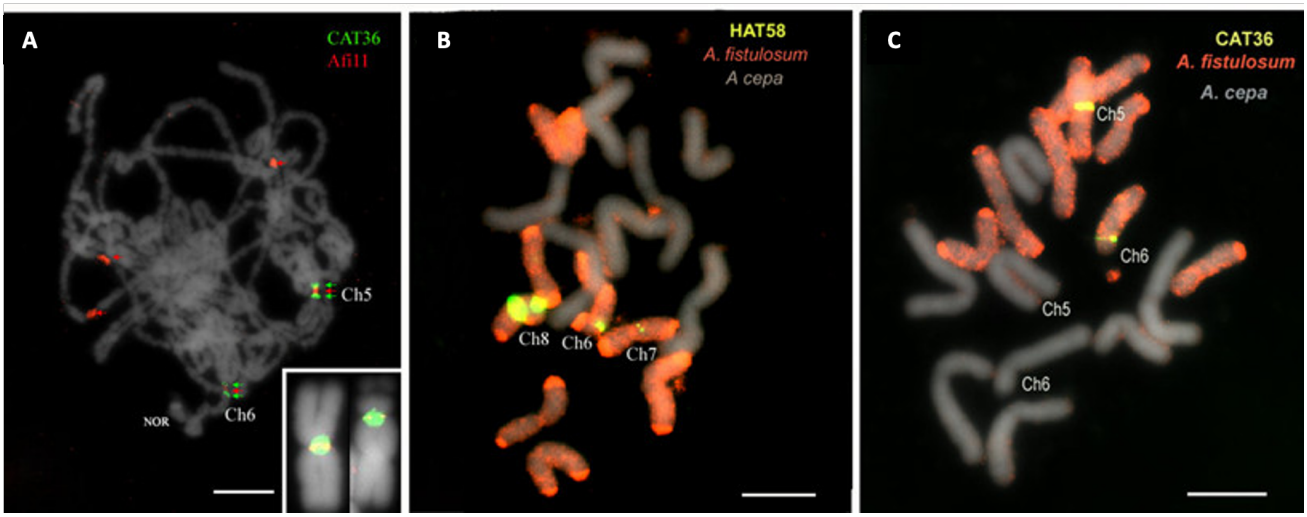

Рисунок 3. (A) FISH с повторами CAT36 и Afi11 на пахитенной хромосоме *A. fistulosum*. На вставке показана совместная локализация CAT36 и Afi11 на митотических метафазных хромосомах 5 (слева) и 6 (справа) *A. fistulosum*. (B) Совместные GISH и FISH с HAT58 на митотических хромосомах *A. × wakegi*. (C) Совместные GISH и FISH с CAT36 на митотических хромосомах *A. × wakegi*. Линейка - 10 мкм.

FISH-анализ двух близкородственных видов показал, что HAT58 и CAT36 видоспецифичны и дают специфические паттерны гибридизации только на хромосомах *A. fistulosum*. Анализ хромосомной локализации этих полиморфных участков был выполнен также у *A. × wakegi* (2n = 2x = 16), природного аллодиплоидного гибрида между *A. cepa* и *A. fistulosum*, обладающего восемью хромосомами *A. cepa* и восемью хромосомами *A.fistulosum* (Рисунок 3B,C). Расположение HAT58 и CAT36 и ранее известных (45S рДНК, 5S рДНК;) ТП, а также субтеломерного повтора длиной 380 п.н. и центромерного повтора на хромосомах *A. fistulosum*, представлены на Рисунке 4.

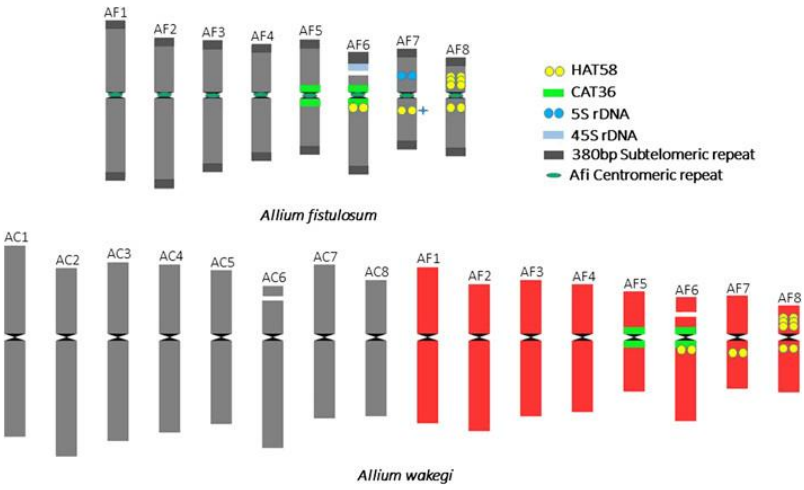

Рисунок 4. Идиограммы хромосом *A. fistulosum* (вверху) и хромосом *A. wakegi* (внизу) с отмеченной локализацией HAT58 и CAT36. Для *A. fistulosum* показаны локализация генов 45S и 5S рРНК, центромерный повтор Afi и субтеломерный тандемный повтор длиной 380 п.н.. Полиморфный сайт HAT58 на хромосоме 7 отмечен звездочкой.

4.1.5 Молекулярно-цитогенетическая характеристика функциональной центромеры *A. fistulosum*

Несмотря на то, что некоторые виды *Allium* являются хорошо известными модельными растениями в цитологии, имеется лишь ограниченная информация об организации функциональных центромер их хромосом. Целью данной части работы было провести молекулярно-цитогенетическое и биоинформатическое исследование CENH3-ассоциированных центромер у *A. fistulosum* и *A. cepa*.

Поиски в базах данных последовательностей схожих с ранее найденным фрагментом центромерного повтора *A. fistulosum* Afi у других организмов не дали результата. Затем была проведена ПЦР с праймерами на фрагмент Afi на геномной ДНК *A. fistulosum*. Это привело к амплификации одного небольшого фрагмента ожидаемой длины (111 п.н.), соответствующего Afi11, и более длинного ПЦР-фрагмента размером около 1 т.п.н..

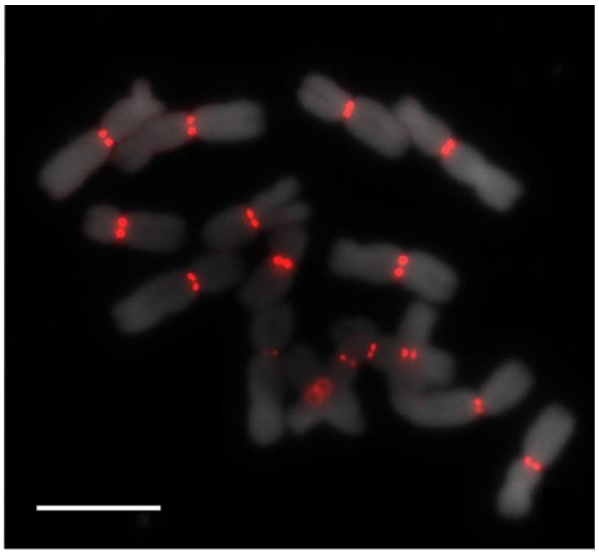

Рисунок 5. Идентификация полноразмерной последовательности центромерных повторов *A. fistulosum*. FISH на митотических метафазных хромосомах *A. fistulosum* с меченым продуктом ПЦР с праймерами Afi11 (B) и с индивидуальной плазмидой, несущей полноразмерный AfCen1K (полоса = 10 мкм).

Продукт ПЦР клонировали в плазмидный вектор и секвенировали. FISH с отдельными клонами, несущими вставки размером в тысячу оснований, приводила к сигналам, расположенным только на центромерах всех хромосом *A. fistulosum* (Рисунок 5). Этот повтор был назван «AfCen1K» (номер GenBank: MT374062). Используя последовательность AfCen1K, был проведён поиск схожих последовательностей в репитоме и сателлитоме *A. cepa*. Проведённые сравнительный молекулярно-цитогенетический и биоинформатический анализы центромеры двух видов *Allium* показали, что центромеры двух видов содержат повторяющиеся последовательности с частичным сходством, но различаются по структуре последовательностей и хромосомной организации.

Чтобы более детально изучить организацию центромер лука, был проведён анализ парных ридов генома. Результаты этого эксперимента и FISH с мечеными клонами BAC, содержащими вставку ДНК хлоропластов, продемонстрировали множественные вставки хлоропластной ДНК в ядерный геном у обоих видов, при этом некоторые вставки были локализованы в центромерных областях. Полученные результаты впервые показывают общую картину организации центромеры *A. fistulosum*.

Чтобы найти новые тандемные повторы, расположенные в центромерных областях *A. cepa* и *A. fistulosum*, мы провели сравнительный анализ повторяющихся последовательностей обоих видов с помощью инструментов RepeatExplorer2 и TAREAN. Затем мы классифицировали повторяющиеся последовательности в соответствии с гомологией их последовательностей с известными семействами повторов (например, мобильными элементами) и особенностями организации генома (тандемные повторы). Один кластер (CL137), идентифицированный с помощью программного обеспечения TAREAN и соответствующий тандемному повтору длиной 276 п.н. (TR2CL137), не показал сходства с известными тандемными повторами *Allium*. Количество копий этого повтора в геноме *A. cepa* составило 160 000 копий/2C (0,14% генома). Чтобы изучить хромосомную организацию нового тандемного повтора TR2CL137, была проведена FISH с этим повтором на хромосомах *A. cepa*. Это продемонстрировало, что TR2CL137 имеет (пери)центромерную локализацию на хромосоме 6. Эта хромосома отличается от других хромосом в кариотипе *A. cepa* NOR-районом на коротком плече. Таким образом, был идентифицирован новый тандемный повтор *A. cepa* с (пери)центромерной локализацией на хромосоме 6.

Мы экспериментально проверили результаты биоинформатического анализа транскрипции с помощью ОТ-ПЦР. Для этого была проведена ОТ-ПЦР с суммарной РНК (поли-А + и поли-А - РНК) *A. cepa* и *A. fistulosum*. Этот эксперимент выявил несколько продуктов ОТ-ПЦР для *A. cepa* и *A. fistulosum*. Мы также предсказали мотивы TSS и TATA-box в CTR *Allium* с использованием программного обеспечения TSSPlant и plantCARE (только для TATA-box). Результаты показали несколько автономных последовательностей TATA-box и одну комбинацию предсказанных сайтов TATA-box и TSS в положении 1159 п.н. (AfCen1K, оценка TATA-box = 8,1926; оценка TSS > 1,9) и 1011 п.н. (AcCen1K, оценка TATA-box = 4,2; оценка TSS > 1,9). Таким образом, на основании этих результатов можно сделать вывод, что повторы AfCen1K и AcCen1K транскрибируются у *A. fistulosum* и *A. cepa*, но транскрипты не полиаденилированы.

В заключение, полученные результаты показали, что центромеры *A. cepa* и *A. fistulosum* содержат длинные (~1,25 т.п.о.) тандемные повторы (AcCen1K и AfCen1K). Центромеры двух видов *Allium* содержат повторяющиеся последовательности с частичным сходством, но различаются хромосомной и геномной организацией, а также структурой последовательностей. Повторы AcCen1K и AfCen1K транскрибируются и их транскрипты не полиаденилированы. Области центромер этих видов содержат вставки ретротранспозонов и ДНК органелл. В рамках работы идентифицирован новый тандемный повтор перицентромерной области хромосом *A. cepa*.

4.1.6 Идентификация тандемных повторов *Rosa wichurana* и *Rosa chinensis* для цитогенетического маркирования хромосом и аннотации генома

Род *Rosa* принадлежит семейству розоцветные, состоит примерно из 200 видов и 20 000 культурных сортов, большинство из которых имеют сложное

гибридное происхождение. Некоторые характеристики розы делают ее возможным кандидатом в качестве модельного организма для геномных исследований древесных видов. Проведение цитогенетического анализа роз затруднено из-за очень маленьких и трудно отличимых хромосом. Кроме этого, для роз характерен низкий митотический индекс и слабое развитие корней. Поэтому поиск тандемных повторов, как цитогенетических маркеров, это актуальная задача для видов роз.

Репитом видов роз не был изучен методами *de novo* аннотирования. Для определения состава репитома четырёх видов роз (*R. foetida, R. gallica, R. rugosa*, *R. wichurana*) было проведено секвенирование на платформе Illumina. Анализ результатов показал, что от 32 до 39% генома приходится на высоко- и среднеповторяющиеся последовательности. Дальнейшая аннотация репитома позволила установить, что геномы всех четырёх видов содержат два типа тандемных повторов с длиной мономера 130 - 160 п.н. и сходством между видами на уровне 85%.

Далее был проведён более детальный анализ тандемных повторов вида *R. wichurana*. По результатам кластерного анализа и аннотации репитома были выделены два кластера, CL8 и CL24, содержащие геномные риды предполагаемых тандемных повторов. Для определения хромосомной локализации найденных повторов была проведена FISH с мечеными олигозондами и митотическими хромосомами *R. wichurana*. Результаты FISH показали, что повтор CL8 локализуется в центромерной области пяти хромосом: хромосомы 1, 2, 4, 5 и 7. Повтор CL24 локализуется в центромерной области трёх хромосом 3, 6 и 7 (Рисунок 6). Для определения более точной хромосомной локализации была проведена FISH на пахитенных хромосомах *R. wichurana*. Результаты этого эксперимента также свидетельствуют, что оба повтора локализованы в самой центромерной области

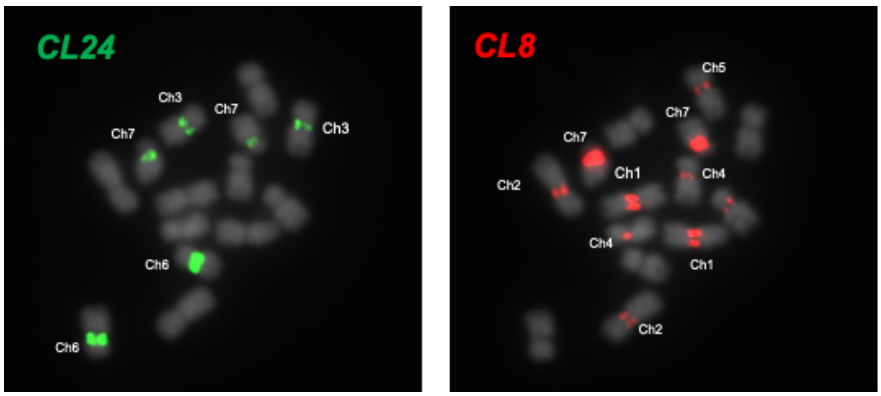

Рисунок 6. FISH с мечеными зондами на тандемные повторы CL8 и CL24 и митотическими метафазными хромосомами *R. wichurana*.

Дифференциальное хромосомное положение идентифицированных повторов делает их кандидатами в цитогенетические маркеры. Впервые проведён анализ репитома видов роз (*R. wichurana, R. gallica, R. rugosa*, *R. foetida*) и найдены высококопийные тандемные повторы с центромерной локализацией на хромосомах *R. wichurana*. Совместное использование проб на найденные повторы и теломерный повтор *Arabidopsis*-типа позволяют проводить идентификацию хромосом *R. wichurana*.

Rosa chinensis – это один из десятка видов роз, которые были использованы для получения сотен современных культурных сортов. Для ускорения селекционного процесса и поиска генов, контролирующих важные признаки, необходимо иметь референсный геном. Поэтому была проведена сборка и аннотация генома *R. chinensis* с использованием гаплоидной линии розы, полученной из старого китайского сорта Old Blush. Сборка генома была проведена с использованием ридов Illumina и PacBio, полученных с геномной ДНК, выделенной из удвоенного гаплоида HapOB. Данные секвенирования PacBio были собраны с помощью CANU, в результате чего был получен 551 контиг (N50 3,4 Mb), что составляет общую длину 512 Mb.

Поиск центромерных регионов проводили с помощью программы RepeatExplorer и данных Illumina. Был обнаружен новый высокопредставленный тандемный повтор OBC226 (центромерный повтор «Old Blush»; Рисунок 7A). Данный повтор занимает 0,06% генома с более чем 2000 копий на гаплоидный геном и имеет длину 159 п.н.. ПЦР подтвердила тандемную организацию этого повтора (Рисунок 7B). FISH-анализ однозначно показал локализацию повтора в центромерных районах четырех из семи хромосом: Chr2, Chr5, Chr6 и Chr7 (Рисунок 7C). Картирование последовательности повторов OBC226 выявило области с высоким покрытием на всех псевдохромосомах HapOB, кроме Chr1, что объясняет, почему нет четких FISH сигналов на этой хромосоме (Рисунок 7D). На Chr3 и Chr4 количество копий OBC226, вероятно, было слишком низким, чтобы его можно было обнаружить с помощью FISH.

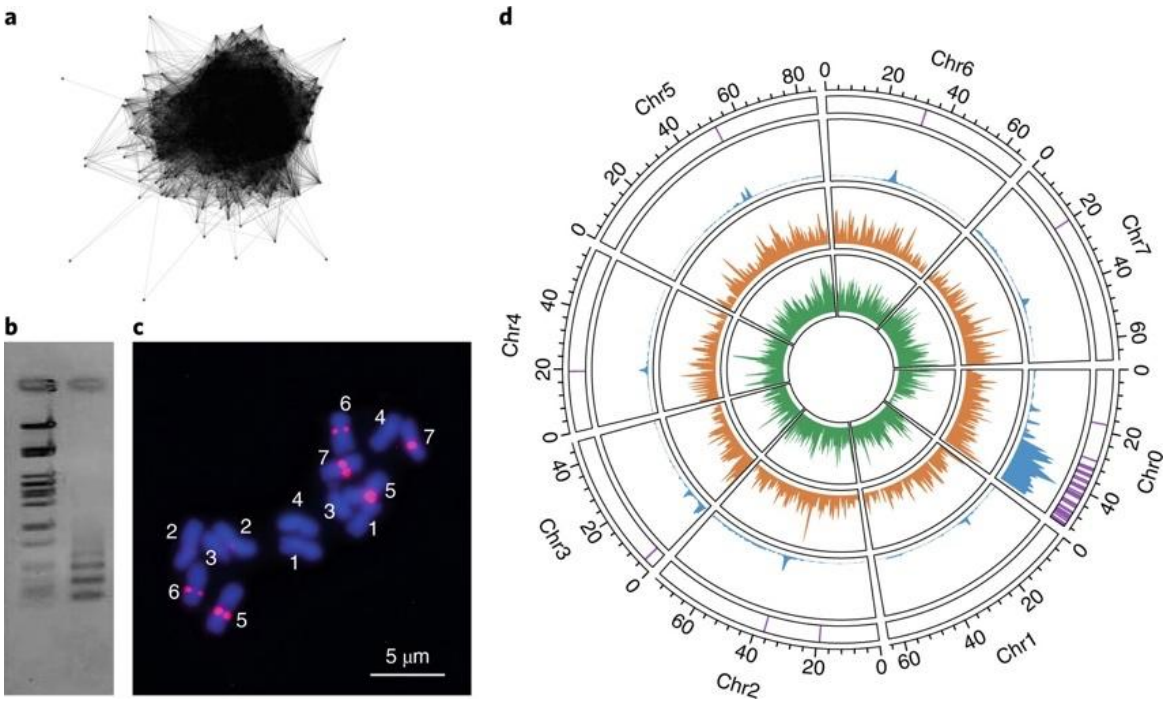

Рисунок 7. Идентификация центромерных областей в референсном геноме HapOB. (a) Кластер CL226. (b) электрофорез фрагментов тандемных повторов, амплифицированных с использованием праймеров для ПЦР OBC226 (c) FISH с олигозондами OBC226 на метафазных хромосомах *R. chinensis*. (d) Circos распределения OBC226 (фиолетовый), прицентромерной области (синий),

повторяющихся элементов *Ty3/Gypsy* (оранжевый) и повторяющихся элементов *Ty1/Copia* (зеленый) вдоль семи псевдохромосом и Chr0 (шкала в Mb).

Эти результаты подтверждают положение центромерных областей на семи псевдохромосомах и показывают высокое содержание повторяющихся последовательностей и низкое содержание генов в так называемой Chr0.

4.1.7 Анализ тандемных повторов в геноме мха, *Physcomitrella patens*

Physcomitrium (Physcomitrella) patens (Hedwig, 1801) Bruch & Schimper, 1849 – это широко используемое модельное растение для исследований в области молекулярной биологии и биологии развития, эволюции и биохимии. Промежуточное филогенетическое положение мхов между зелеными водорослями и цветковыми растениями делает это растение уникальным для эволюционных исследований. К моменту начала исследований была проведена сборка генома мха на уровне хромосом, и доступны различные наборы транскриптомных, эпигенетических, протеомных, пептидомных данных, а также соответствующие инструменты. Однако молекулярно-цитогенетическое изучение *P. patens* не проводилось. Поэтому в данной работе были проведены исследования, направленные на поиск и хромосомную характеристику новых хромосомных маркеров *P. patens*.

Согласно результатам pyTanFinder, 7 (37%) ТП имеют высокую (> 18 000 п.н., hcTRs) и 12 (63%) ТП имеют низкую (<15 000 п.н., lcTRs) общую копийность. Нам удалось разработать праймеры для 5 hcTR и получить продукты ПЦР в виде лестницы или шмера, которые являются известными характеристиками ТП. Для определения хромосомной организации найденных ТП была использована FISH. Молекулярно-цитогенетический подход к визуализации локусов последовательностей ДНК на хромосомах и ядрах является сложным для бриофитов. Для проведения пилотного эксперимента FISH был оптимизирован протокол «SteamDrop» для приготовления препаратов хромосом мха. Были использованы различные типы материала, включая протопласт, протонемы и несозревший спорофит. Были разработаны 19 олигонуклеотидных зондов TAMRA для проведения анализа FISH. Чтобы подтвердить, что полученные слайды подходят для экспериментов FISH, были использованы известные тандемно организованные последовательности, теломерный повтор арабидопсиса ((TTTAGGG)n) и рДНК 45S в качестве положительных контролей. Эксперименты FISH выявили много точечных и несколько отчетливых сигналов для зондов теломер и 45S рДНК, соответственно, что позволяет предположить, что препараты подходят для анализа FISH хромосом мха.

Затем были проведены эксперименты FISH на ядрах для 19 ТП. Эти эксперименты выявили 5 ТП, для которых FISH-сигналы были обнаружены на ядрах. Три повтора (Pp602_86, Pp21_215, Pp592_108) давали несколько сигналов, которые занимали две отдельные территории в ядре. Сигналы FISH от одного TR, Pp19_95 (размер мономера 95 п.н.), были связаны с областями гетерохроматина ядра, обнаруженными с помощью DAPI окрашивания. Сигналы FISH от другого ТП, Pp20_76, были локализованы в одной ядерной области, которая находилась в

непосредственной близости от ядрышка (перинуклеолярная область), которую можно хорошо различить при окрашивании DAPI. В отличие от ТП Pp19_95, профиль DAPI из локусов гибридизации Pp20_76 не обнаруживает явных отличий от соседних ядерных областей. Более пристальный взгляд на сигналы FISH показывает, что локусы Pp20_76 организованы в виде каплевидной структуры. Таким образом, FISH-анализ ядер для 19 ТП, идентифицированных pyTanFinder, показал 5 ТП с ярко выраженными сигналами. Более того, один из повторов (Pp19_95) был связан со структурами гетерохроматина, а другой (Pp20_76) — с перинуклеолярными тельцами.

Из-за особого расположения Pp20_76 в ядре (около ядрышка) и обнаруженных ядерных телец, обогащенных этим ТП, этот ТП был назван PpNATR76 (76 п.н. *P. patens* periNucleolar Associated Tandem Repeat) и проанализирован более детально. Выравнивание 200 последовательностей PpNATR76, обнаруженных в геноме мха, показало высокий уровень консервативности между мономерами. Кроме того, анализ последовательности консенсусного мономера PpNATR76 выявил длинную полипиримидиновую последовательность (мотив (CCT)n). Чтобы определить, почему ДНК PpNATR76 располагалась проксимальнее ядрышка, были сопоставлены рДНК 45S с геномом мха. Хромосомное расположение 45S рДНК и PpNATR76 было идентичным на хромосомах 20 и 26, где они занимали 250 т.п.н. и 16 т.п.н., соответственно. Более того, подробный анализ локусов показал, что PpNATR76 располагался между генами 45S рДНК, в областях IGS. Были также проверены данные RNAseq и обнаружен высокий уровень покрытия прочтений RNAseq для этой области, как и ожидалось для локусов рДНК. Из-за транскрипционной активности области, занимающей PpNATR7, следующей целью было найти транскрипты *P. patens*, содержащие ТП PpNATR76. Этот анализ выявил 16 транскриптов, гены которых располагались на 5 хромосомах (Chr20, Chr19, Chr4, Chr17, Chr14). Только 4 транскрипта содержали аннотированные канонические ORF (Pp3c19_9270V3.1, p3c19_9271V3.1, Pp3c4_8299V3.1 и Pp3c14_12290V3.1). В совокупности эти данные доказывают существование транскриптов pPNATR76 в соматических клетках и убедительно свидетельствуют о том, что PpNATR76 транскрибируется как часть белок-кодирующих, так и некодирующих (днРНК) РНК.

В заключение, ТП с разным размером мономеров являются неотъемлемой частью большинства эукариотических организмов, в которых они участвуют в разнообразных биологических процессах. Хотя было предпринято много усилий, чтобы понять геномную организацию, структуру и эволюцию ТП, их функции в клетке все еще плохо изучены. В рамках работы определены 19 ТП, из которых 5 ТП генерировали сигналы FISH. Были обнаружены как гетерохроматин-ассоциированные, так и транскрибируемые ТП. Геномный и транскриптомный анализы идентифицировали IGS-ассоциированный ТП, PpNATR76, который был изолирован в перинуклеолярном пространстве и транскрибировался как часть днРНК.

4.2 Транскриптомные особенности ретротранспозонов растений

Транскрипция ретротранспозонов (RTE) игнорируется во время аннотации генома, т.к. RNAseq риды от RTE часто не могут быть отнесены к одному участку

генома после картирования. Поэтому наши знания об особенностях организации и экспрессии RTE растений крайне скудны. В этой части работы был проведён полногеномный анализ транскрипции RTE у разных видов растений, используя данные Illumina и нанопорового секвенирования.

4.2.1 Десятки ретротранспозонов экспрессируются в геноме подсолнечника (*Helianthus annuus*)

Для выявления LTR-ретротранспозонов (RTE) с высоким уровнем экспрессии был разработан подход, представленный на Рисунке 8A. Были использованы риды RNAseq из предыдущих исследований, в которых брали образцы различных тканей и изучали действие стрессовых условий: обработка NaCl (3 и 12 ч), обработка полиэтиленгликолем (ПЭГ) (6 и 12 ч) и обработка гормонами (обработка метилжасмонатом и абсцизовой кислотой (АБК)).

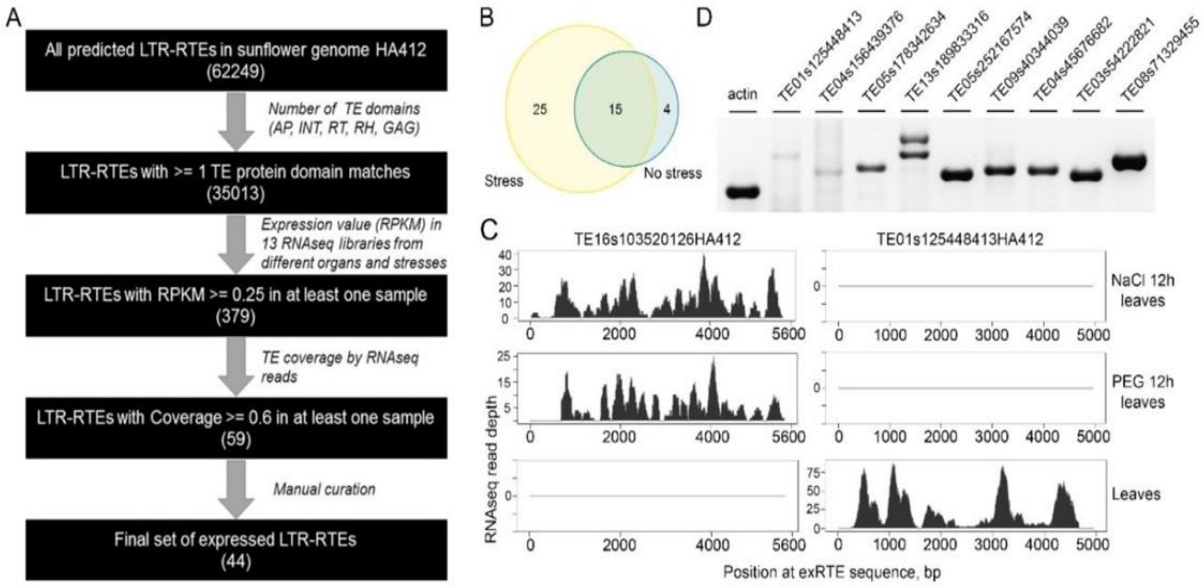

Рисунок 8. RNAseq идентификация экспрессирующихся LTR ретротранспозонов (exRTE) подсолнечника. (A) Схема подхода, используемого в этом исследовании. (B) диаграмма Венна, показывающая количество exRTE с обнаруживаемой экспрессией в стрессовых и нестрессовых условиях. (C) Примеры графиков покрытия RNAseq для двух exRTE с различными паттернами экспрессии. (D) ОТ-ПЦР, показывающая экспрессию выбранных exRTE в пятидневных проростках.

Используя данный подход, был получен набор из 44 уникальных и сильно экспрессирующихся RTE (exRTE). Из них 19 exRTE (42%) и 25 exRTE экспрессировались в нестрессовых и стрессовых условиях (Рисунок 8B). Некоторые RTE продемонстрировали ткане- и стресс-специфические паттерны экспрессии (Рисунок 8C). Например, exRTE (TE01s125448413HA412) надсемейства *Ty3/Gypsy* транскрибировался почти исключительно в листьях и тычинках, не подвергшихся стрессу, а exRTE (TE16s103520126HA412) надсемейства *Ty1/Copia* транскрибировался только в листьях, подвергшихся осмотическому стрессу. Результаты были проверены с помощью ОТ-ПЦР с праймерами, разработанными для десяти случайно выбранных RTE, и для девяти (90%) из них были получены продукты амплификации (Рисунок 8D). Таким

образом, были идентифицированы десятки RTE подсолнечника, экспрессирующихся в стрессовых и нестрессовых условиях.

Чтобы понять особенности exRTE и найти отличия от неэкспрессирующихся RTE (n-exRTE), был проведен всесторонний биоинформатический анализ по разным критериям. Было обнаружено значительное обогащение exRTE, экспрессирующихся как в стрессовых, так и в нестрессовых условиях, элементами из подсемейства *Ty1/Copia* (точный критерий Фишера, значение $p < 0,01$) по сравнению с n-exRTE. Было обнаружено, что время инсерции в геном элементов exRTE и n-exRTE Ivana и Tork существенно отличается. Кроме того, распределение времени инсерции всех RTE в геноме подсолнечника было бимодальным: <0,5 млн лет назад (недавние инсерции) и >0,5 млн лет назад (поздние инсерции). Вместе эти результаты показали, что exRTE в основном представляют малокопийные элементы суперсемейства *Ty1/Copia*, недавно интегрированные в геном.

Также exRTE имели более длинные ОРС (критерий суммы рангов Уилкоксона с коррекцией непрерывности, p-значение = $6,759 \times 10^{-5}$) и существенно отличались от n-exRTE в отношении набора кодируемых белков. Обе группы содержали схожую долю RTE, кодирующих белки RT, РНКазы H и INT. Однако доля ОРС, кодирующих GAG (23 exRTE) и AP (23 exRTE), была значительно выше в exRTE (точный критерий Фишера для данных подсчета, p-значения составляли $1,138 \times 10^{-11}$ и 0,0002487 для GAG и AP, соответственно), чем в n-exRTE. Кроме того, другие комбинации белков RTE, содержащие GAG, были представлены в exRTE, что еще больше подчеркивает белок GAG как отличительную особенность exRTE.

Близость к соседним генам считается важным фактором экспрессии RTE. Анализ расстояния до генов exRTE и n-exRTE показал, что exRTE, как правило, были значительно ближе к аннотированным генам, чем n-exRTE (критерий суммы рангов Уилкоксона, значение $p = 1,065 \times 10^{-12}$) (Рисунок 9).

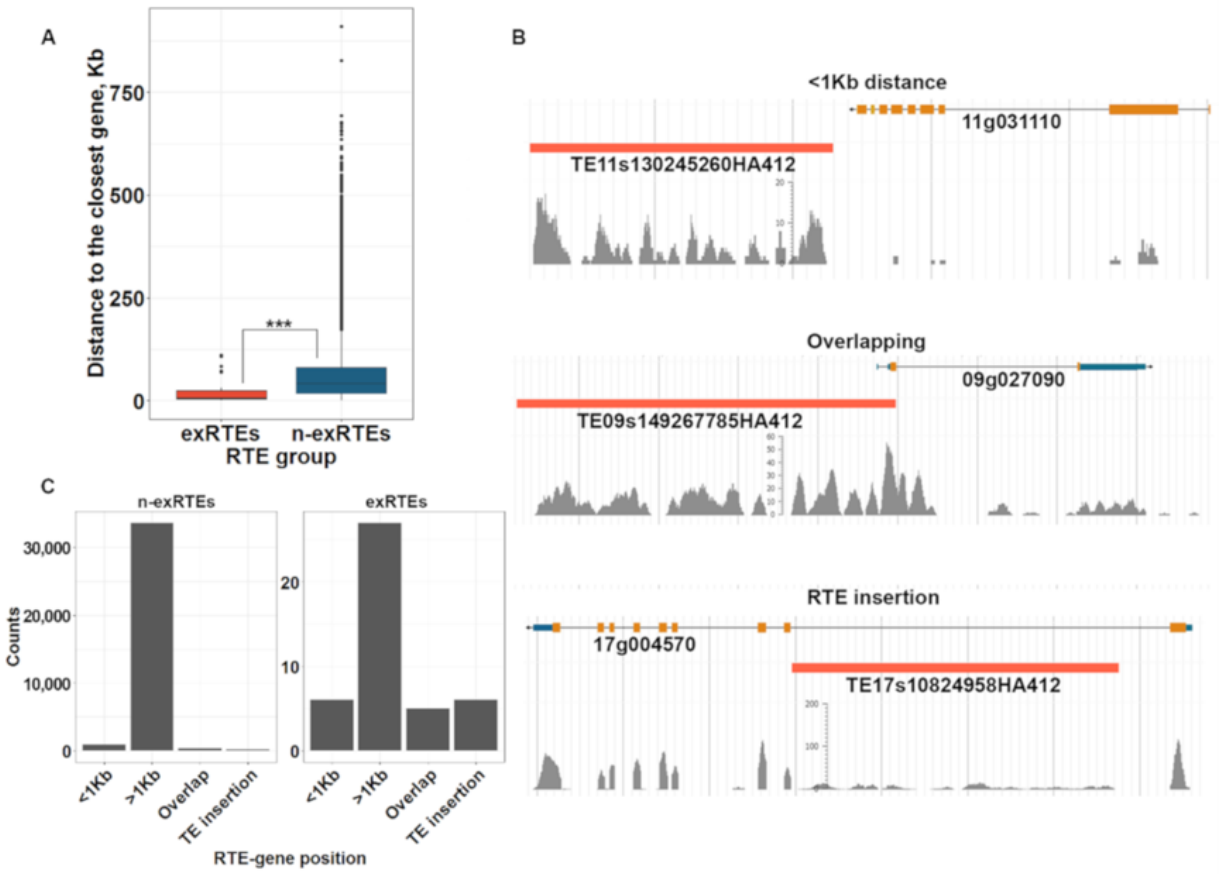

Рисунок 9. Расстояние от (n-) exRTE до соседних генов. (A): Боксплот расстояния (т.п.н.) до ближайших генов для exRTE и n-exRTE. Три звездочки указывают на значительные различия, основанные на p-значении критерия суммы рангов Уилкоксона <0,001. (B) Примеры exRTE и их близость к генам. (C) Гистограмма, показывающая количество (n-)exRTE с разным расстоянием до ближайших генов.

Все RTE были классифицированы по расстоянию до ближайших генов (Рисунок 9A): (1) удаленно расположенные (расстояние >1 т.п.о. между RTE и геном); (2) близко расположенные (расстояние <1 т.п.н. между RTE и геном); (3) перекрытие и (4) инсерция RTE (RTE находится внутри аннотированного гена; Рисунок 9B). 27 exRTE были расположены > 1000 п.н. от ближайшего гена и 17 exRTE (38,6%) были распределены среди трех других категорий (Рисунок 9A). Таким образом, на экспрессию лишь нескольких exRTE могла влиять их близость к гену.

Жизненный цикл RTE может зависеть от процессинга РНК, включая сплайсинг. Чтобы получить последовательности отдельных транскриптов RTE, было проведено прямое нанопоровое секвенирование (DRS) поли-A+ РНК пятидневных проростков подсолнечника и получено 380 000 DRS ридов. Картирование ридов на геном HA412 подсолнечника показали экспрессию трех RTE, включая два exRTE (TE01s34770891HA412 и TE08s32407041HA412) и один n-exRTE (TE11s205313630HA412). Из них наибольшее число ридов было получено для двух полноразмерных RTE суперсемейства *Ty1/Copia*, TE08s32407041HA412 (названный «Tyran») и TE11s205313630HA412 (названный

«Varan»). Их экспрессия была проверена с помощью ОТ-ПЦР. Экспрессия *Tyran* была обнаружена во всех экспериментах с RNAseq.

В заключение, проведённый полногеномный анализ позволил идентифицировать 44 ретротранспозона со значительным уровнем экспрессии в стрессовых и/или нестрессовых условиях. Было показано, что геномная организация exRTE отличается от неэкспрессирующихся RTE, включая недавнее время встраивания exRTE, близость к генам, низкое число копий и обогащение открытыми рамками считывания (ORF), кодирующими GAG с одним РНК-связывающим доменом.

4.2.2 Экспрессия LTR ретротранспозонов подсолнечника под действием эпигенетического стресса

Для более детального изучения экспрессии ретротранспозонов подсолнечника был использован новый подход, при котором ретротранскриптом изучался при помощи нанопорового секвенирования кДНК растений, находящихся под действием эпигенетического стресса (аманитина и зебуларина, AZ, TEgenesis) и теплового стресса (37C).

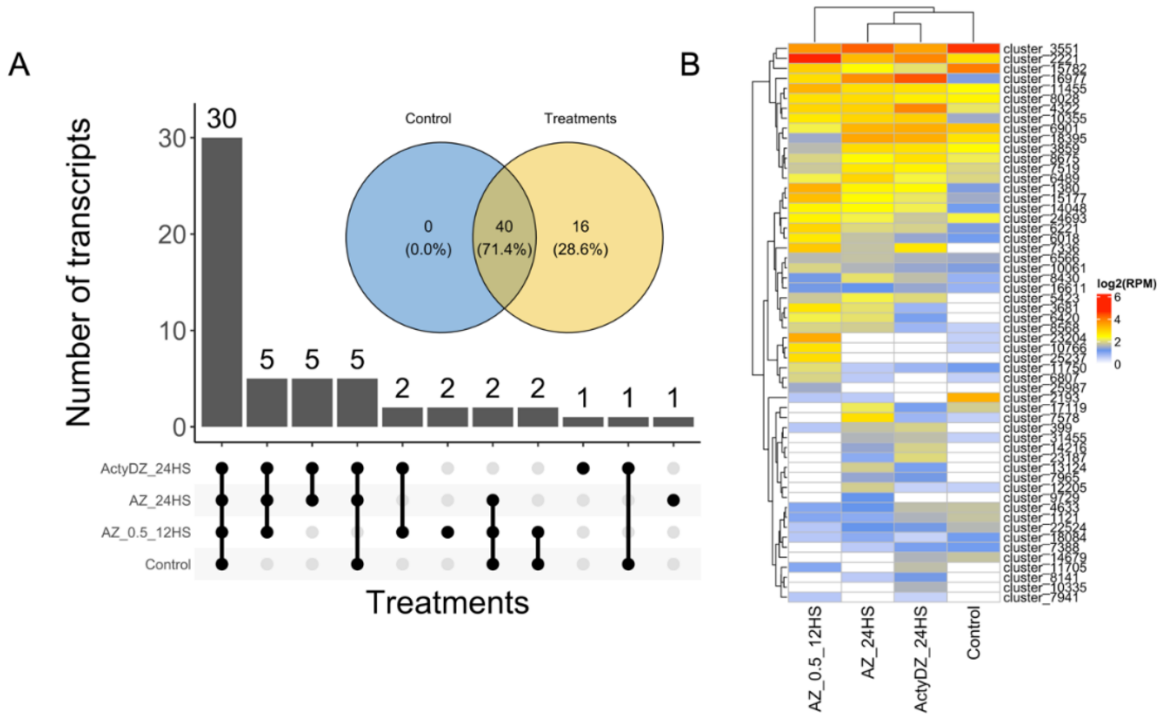

Рисунок 10. Транскрипты RTE, идентифицированные с помощью нанопорового секвенирования кДНК. (A) График и диаграмма Венна, показывающая количество транскриптов, обнаруженных в различных экспериментальных условиях. (B) Тепловая карта значений экспрессии (log2(RPM)) транскриптов RTE в четырех экспериментальных условиях: контроль (нормальная среда MS без теплового стресса), AZ_0,5_12HS (2 мкг/мл альфа-аманитина, 4 мкг/мл зебуларина, 12 ч теплового стресса (37 °C)), AZ_24HS (4 мкг/мл альфа-аманитина, 8 мкг/мл зебуларина, 24 ч теплового стресса (37 °C), ActyDZ_24HS (4 мкг/мл актиномицина D, 8 мкг /мл зебуларина, 24 ч теплового стресса (37 °C)).

Для создания эпигенетического стресса подсолнечника мы проращивали семена на среде MS, содержащей комбинацию зебуларина (Z, блокирует метилирование ДНК) с одним из ингибиторов полимеразы II — альфа-аманитином (A) или актиномицином D (ActyD) (Pol II играет ключевую роль в РНК-зависимом метилировании ДНК). В общей сложности были идентифицированы 56 локусов генома подсолнечника, из которых 40 RTE экспрессировались в контроле и в любом из образцов (Рисунок 10A,B). Проведённые структурный и белковый анализ выявили три категории локусов, из которых 34 (60%) транскрипта кодируются потенциально автономными (LTR-RTE) и неавтономными (TR-RTE) RTE и 40% локусов не имеют своих LTR-последовательностей и, следовательно, эти транскрипты (nonLTR-RTE) напоминают обычные эукариотические гены, кодирующие RTE-родственные белки с неизвестными функциями.

4.2.3 Ретротранскриптом развивающейся зерновки тритикале

Транскрипция ретротранспозонов в вегетативных частях растений, как это было показано для подсолнечника и других видов растений, не обеспечивает абсолютную возможность передачи новых копий следующему поколению при семенном размножении. Поэтому особый интерес для изучения МЭ, как внутренних мутагенов, представляют генеративные органы растений и стадии развития, которые непосредственно связаны с размножением растений, как, например, зерновка злаковых. Для выявления таких экспрессирующихся и мобильно активных RTE был проведён детальный анализ транскриптома развивающейся зерновки тритикале с использованием прямого нанопоровое секвенирование РНК и кДНК. Поэтому мы проанализировали собранные транскрипты по наличию открытых рамок считывания, кодирующих белки, родственные МЭ. Мы не обнаружили транскриптов, кодирующих полный набор белков МЭ, но мы идентифицировали 20 транскриптов (RTE-РНК), несущих одну ОРС со сходством к различным белкам LTR ретротранспозонов. Пять и десять RTE-РНК (15) были транскрибированы с полноразмерных или неполных RTE, соответственно. Таким образом, было обнаружено, что почти 25% RTE-РНК транскрибируются с потенциально автономных RTE. Было обнаружено, что большинство RTE-РНК несут ОРС (75%, 15), кодирующую один белок GAG (GAG-РНК). Из них 8 и 5 RTE-РНК кодировались полноразмерными и неполными копиями LTR ретротранспозонов, соответственно. Анализ ридов показал, что сплайсинг изоформы GAG-РНК имеет решающее значение для трансляции белка GAG.

Затем мы оценили паттерны экспрессии GAG-РНК на нескольких стадиях развития зерновки, а также тканях листа и пестиках пшеницы. Девять из пятнадцати локусов GAG -РНК имели специфический паттерн экспрессии с максимальным уровнем экспрессии на ранних стадиях развития или в пестиках. Таким образом, данные экспрессии показали, что большинство идентифицированных GAG-РНК тритикале также экспрессируются во время развития зерновки пшеницы, а некоторые RTE экспрессируют как полноразмерную РНК, так и сплайсированную РНК (shGAG).

4.3 Активные мобильные элементы растений и особенности их инсерций

4.3.1 nanotei: программа для идентификации нереференсных инсерций транспозонов по данным полногеномного нанопорового секвенирования

Идентификация инсерций мобильных элементов (TEI) на основе данных ONT имеет большую чувствительность, точность и требует меньшего покрытия генома. Однако программы для простого и автоматического поиска TEI и идентификации соответствующих мобильных элементов (МЭ) с использованием данных ONT с низким покрытием не было опубликовано. Чтобы заполнить этот пробел, нами был разработан новый пайплайн, названный nanotei (https://github.com/Kirovez/nanotei), который позволяет идентифицировать TEI по данным нанопорового секвенирования с низким покрытием. nanotei требует четыре входных файла: файл bam после картирования ридов ONT на референсный геном, fasta файл референсной геномной последовательности, fastq файл ридов ONT и bed файл, содержащий координаты аннотированных мобильных элементов в референсном геноме. Принцип nanotei показан на Рисунке 11.

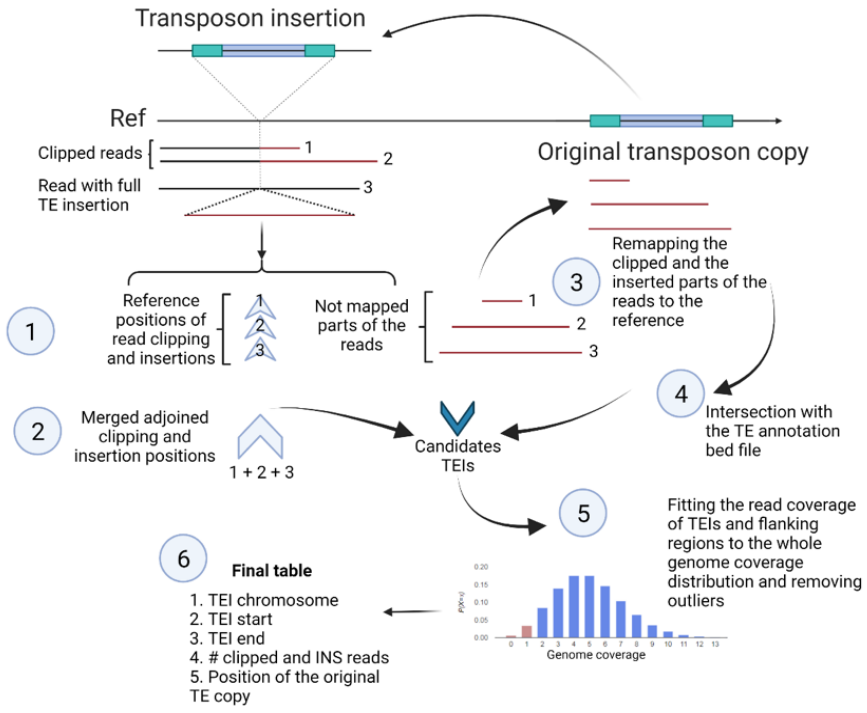

Рисунок 11. Схематическое изображение nanotei. Перечислены основные этапы. Красные участки ридов соответствуют последовательностям, содержащим МЭ.

Итоговая таблица nanotei содержит всю информацию о положении каждой инсерции и о вызвавших их МЭ. Для тестирования nanotei были использованы ONT риды (Col-0) *A. thaliana*. Было выполнено секвенирование двух образцов Col-0 и получено ~60 000 ридов (~7-кратное покрытие генома, N50 ~12 Kb). Анализ с помощью nanotei показал 46 TEI (colTEI) образованных 43 различными копиями

МЭ. Большинство colTEI (44) были общими для двух растений Col-0 (Рисунок 12A).

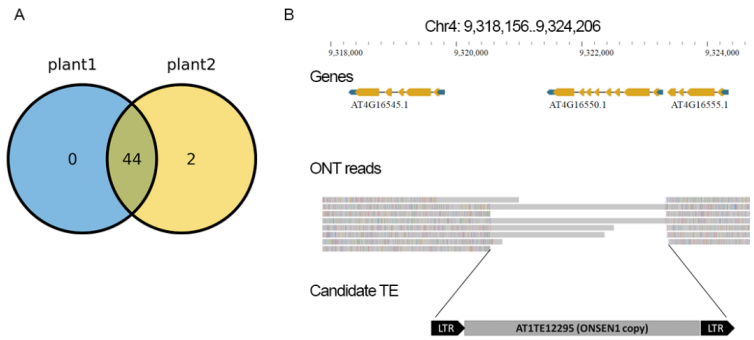

Рисунок 12. (A) Диаграмма Венна, показывающая количество TEI, общих для двух растений Col-0. (B) TEI на хромосоме 4 и схематическое изображение транспозона AT1TE12295, который произвёл инсерцию.

Чтобы убедиться, что обнаруженные colTEI не являются специфическими для наших растений *A. thaliana*, были использованы дополнительные риды ONT из базы данных NCBI (ERR5530736) в качестве контроля. Все colTEI также были идентифицированы и в этом наборе данных. Сборка генома *A. thaliana* (TAIR10) включает не менее 100 аннотированных гэпов, как результат неправильной сборки генома. Было предположено, что colTEI локализуются именно в этих местах генома. Кроме того, такие случаи необходимо отфильтровывать, потому что они будут мешать идентификации реальных TEI в дальнейшем. Чтобы обнаружить такие colTEI, связанные с гэпами сборки генома, было проведено сравнение colTEI с гэпами в геноме TAIR10 (последовательности с 3 или более «N»). Было обнаружено, что только 10 TEI (22%) являются частью гэпов (Рисунок 12B). Эти результаты показывают, что >40 инсерций не отображены в геноме. Было также обнаружено, что элемент ONSEN имеет две тандемно организованные копии на хромосоме 1, но только одна из этих копий отсутствует в сборке TAIR10.

На следующем этапе программ nanotei была использована для детекции новых инсерций активных мобильных элементов *A. thaliana*. Для этого было использовано растение *ddm1 A. thaliana,* содержащее мутацию в гене *DDM1* (снижение метилирования ДНК), что вызвало гипометилирование цитозина во всех контекстах генома. Было проведено полногеномное нанопоровое секвенирование двух растений *ddm1* (G-ddm1-1 и G-ddm1-2). Используя эти данные, nanotei детектировал 38 и 33 TEI в растениях G-ddm1-1 и G-ddm1-2.

В совокупности проведённый анализ мобилома с использованием nanotei и ридов ONT генома *ddm1* позволил одновременно обнаружить все МЭ, активные в *ddm1*, и выявил транспозицию двух новых неавтономных ретротранспозонов, один из которых кодирует полноразмерный белок GAG. Это первое прямое доказательство того, что элементы TR-GAG способны к транспозиции в растениях. Это также подчеркивает, что эволюцию МЭ следует исследовать как сеть функционально связанных автономных и неавтономных элементов.

4.3.2 CANS: Cas9-опосредованное обогащение библиотек для нанопорового секвенирования инсерций транспозонов растений

Идентификация TEI в геноме является сложной задачей, в первую очередь потому, что значительная часть геномов растений состоит из повторяющейся ДНК, несущей множество вставок МЭ. Сочетание избирательного обогащения ДНК фрагментами МЭ с ONT секвенированием значительно улучшает картирование сайтов инсерции. Целью этой части работы была адаптация и оптимизация метода CANS для растений, а также разработка биоинформатического подхода для анализа данных CANS и секвенирования TEI в геноме растений. Наш подробный протокол CANS на protocols.io (https://www.protocols.io/private/DE1CFDE8C8FF11EBA7DA0A58A9FEAC02).

Чтобы продемонстрировать, что CANS подходит для отслеживания инсерций отдельных транспозонов в растениях, был выбран LTR ретротранспозон EVD5 (5333 п.н.), который может генерировать новые инсерции у некоторых мутантов *A. thaliana* (например, *met1* и *ddm1*)). Эксперимент с использованием пяти гРНК произвёл 88000 ридов, при этом было получено 259 (0,3%) целевых ридов после 4 часов секвенирования на проточной ячейке MinION. Оба запуска CANS с использованием геномной ДНК Col-0 привели к 40-кратному покрытию целевых ретротранспозонов EVD5 и EVD1/2 (Рисунок 13). Картированные риды из первого и второго запусков охватывают регион генома длиной 35 000 п.н. и 14 000 п.н. соответственно, включая целевые EVD и фланкирующие области. Широкий охват фланкирующих последовательностей демонстрирует преимущество CANS для идентификации инсерций МЭ в геноме. Проверка расположения гРНК и распределения прочтений показала, что CANS обладает высокой специфичностью, при этом до 82% прочтений картируются на цепях с 3'-концами, незащищенными белком Cas9 и имеющими последовательности PAM (NGG). Это еще одно преимущество CANS, поскольку оно позволяет увеличить охват целевого региона. Также стоит отметить, что основная часть ридов из внутренних частей EVD2 и EVD5 была однозначно отнесена к соответствующей копиям EVD, подтверждая предыдущие результаты о том, что ONT риды кДНК обладают высокой картируемостью.

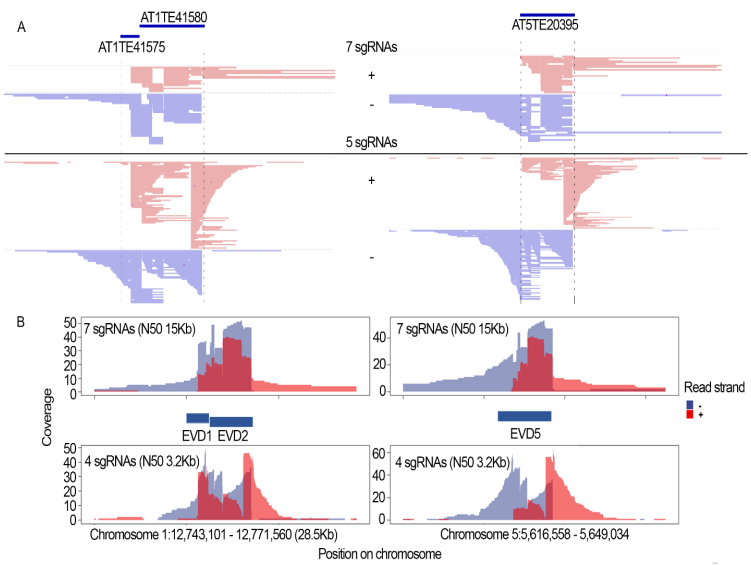

Рисунок 13. Целевое обогащение EVD после CANS с использованием набора из 7 или 5 гРНК. (A) Снимок IGV браузера после выравнивания ридов на копии ретротранспозонов EVD1, EVD2 и EVD5. (B) График покрытия EVD1, EVD2 и EVD5 с использованием прямыми (красный) и обратными (синий) ридами, полученными после CANS с 7 и 5 гРНК.

Чтобы оценить, может ли CANS детектировать новые вставки EVD был использован мутант *A. thaliana ddm1*, у которого активность EVD была описана ранее. В общей сложности было обнаружено 24 инсерции EVD. Из них 19 TEI были детектированы с высокой достоверностью (TEI, поддерживаемые двумя или более ридами). Результаты показывают, что даже низкое покрытие генома (0,2x в нашем эксперименте) может быть достаточным для идентификации всех инсерций Т-ДНК и целевых транспозонов в *A. thaliana* методом CANS, что подчеркивает хорошую чувствительность метода.

4.3.3 NanoCasTE: программа для идентификации нереференсных инсерций транспозонов по данным CANS

Для анализа данных CANS и выявления сайтов инсерций была разработана программа NanoCasTE (https://github.com/Kirovez/NanoCasTE). NanoCasTE включает несколько этапов (Рисунок 14).

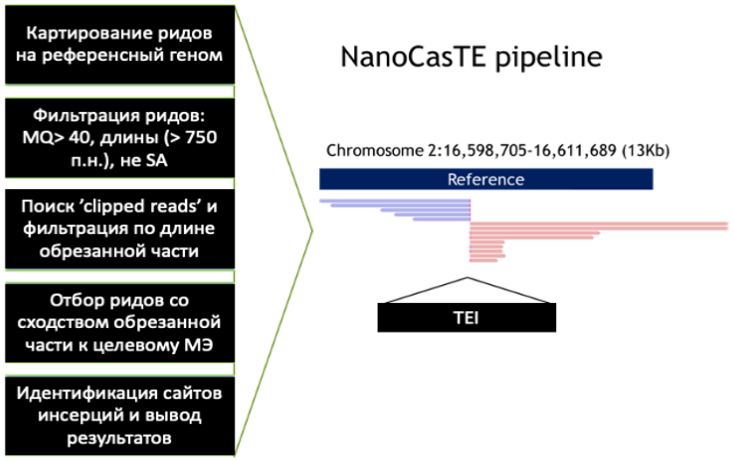

Рисунок 14. Схематическое изображение основных этапов NanoCasTE для идентификации инсерций транспозонов в геноме после CANS.

NanoCasTE использует набор строгих критериев, чтобы специально обнаруживать вставки МЭ в данных CANS и отличать их от шумовых сигналов. NanoCasTE сообщает о предполагаемых координатах инсерций целевого МЭ, а также дополнительную информацию, полезную для дальнейшего анализа, включая количество выбранных прочтений, поддерживающих вставку, общее количество прочтений, охватывающих вставку, цепь, содержащую вставку TE, и длину обрезанных частей выбранных ридов. NanoCasTE был протестирован на данных CANS детекции сайтов инсерции ретротранспозона EVD в геноме *ddm1*.

4.3.4 Геномная организации инсерций LTR ретротранспозона EVD в геноме *ddm1*

Недавние исследования показали, что инсерции некоторых МЭ распределены по геному неслучайно. Но такая картина может быть объяснена давлением отбора. Принимая во внимание, что разработанный метод CANS может детектировать как соматические, так и генетически наследуемые инсерции

мобильных элементов (TEI), это даёт возможность напрямую проверить геномную организацию инсерций МЭ без существенного влияния отбора. Для этого мы выполнили CANS-секвенирование инсерций ретротранспозона EVD в популяции примерно из 50 растений *ddm1*. Используя NanoCasTE, мы обнаружили 851 TEI, в том числе 29 высоконадежных TEI (поддерживаемых двумя или более считываниями) в популяции *ddm1*. Тринадцать TEI были проверены с помощью ПЦР с использованием ДНК индивидуальных растений, в результате каждая инсерция была детектирована в геноме от 1 до 6 растений. Анализ распределения TEI по хромосомам показал, что они, как правило, группировались в прицентромерных областях всех хромосом, тогда как плотность TEI в плечах хромосом была значительно снижена (Рисунок 15).

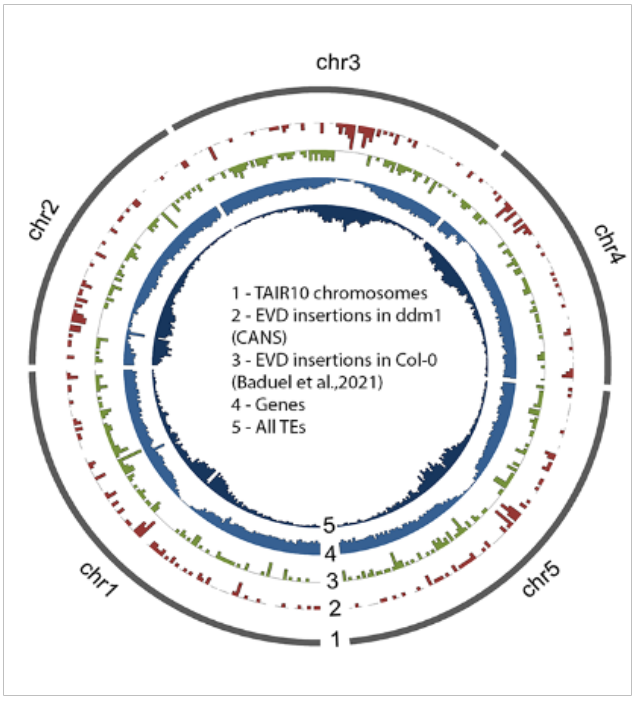

Рисунок 15. Геномная организация инсерций EVD, обнаруженных с помощью CANS в объединенном (~50 растений) образце *ddm1*. Хромосомное распределение инсерций EVD, обнаруженных с помощью CANS в геноме *ddm1* (2) и встречающихся в природных образцах *A.thaliana* (3), аннотированные гены по версии TAIR10 (4) и МЭ (5).

Классификация этих TEI показала, что 54% (462) из них расположены в генных областях, включая 403 (87%) экзонных и 59 (13%) интронных TEI (Рисунок 15). Это значение ожидается случайно, исходя из общей длины генных областей в собранном (119 Mb, TAIR10) геноме *A. thaliana* (68,3 млн.п.н., или 57% секвенированного генома).

Для оценки связи места инсерции с экспрессионной активностью, было проведено прямое секвенирование РНК проростков *ddm1* и обнаружено, что ~ 90% (764) инсерций EVD расположены в областях с низкой экспрессией (<2 ридов) в геноме *ddm1*, что значительно выше, чем ожидалось случайно (p-значение точного критерия Фишера $4{,}25e^{-18}$). Мы также показали значительные различия в частоте вставок EVD в прицентромерных областях *ddm1* и разных экотипов *A. thaliana*. В отличие от *ddm1* в прицентромерных областях разных экотипах *A. thaliana* почти не было обнаружено инсерций EVD (Рисунок 15). Эти результаты впервые доказывают, что в нативной геномной среде EVD имеет тенденцию избегать прицентромерных областей, в то время как мутация *ddm1* смещает плотность инсерций в сторону проксимальных хромосомных областей.

4.3.5 Инсерционный ландшафт ретротранспозона ONSEN, полученный с помощью CANS

Чтобы проверить сделанные выводы, мы использовали ещё один LTR ретротранспозон - ONSEN, транспозиционная активность которого у *A. thaliana* дикого типа может запускаться тепловым стрессом. Мы разработали набор из 6 гРНК, нацеленных на все 8 известных копий ONSEN (ONSEN *1-8*), и провели эксперименты с CANS, как описано выше. Анализ данных с помощью NanoCasTE выявил 519 инсерций. Распределение инсерций ONSEN показало сильное смещение в дистальные регионы, в то время как частота инсерций в области центромер была снижена. Аналогичные результаты были получены при анализе хромосомного распределения инсерций ONSEN, обнаруженных в природных образцах *A. thaliana*. Большинство инсерций ONSEN (83%, 430 из 519) были расположены в генах, что также подтверждается недавними исследованиями (>90% инсерций ONSEN были генными).

4.3.6 Инсерции элементов ONSEN преимущественно возникают в генах с пониженной экспрессией в ответ на тепловой стресс

Инсерции EVD и ONSEN связаны с вариантом гистона H2A.Z, который часто обнаруживается в генах, экспрессия которых изменяется в ответ на стресс. Однако корреляция между уровнем экспрессии генов во время теплового стресса и частотой возникновения инсерций ONSEN в эти гены не изучены. Чтобы прояснить этот момент, мы провели RNAseq анализ растений *Arabidopsis* до (0-часовые образцы) и через 8, 16 или 24 часа после теплового стресса. Проведённый эксперимент и биоинформатический анализ дифференциально-экспрессирующихся генов показывают, что гены с пониженной экспрессией во время теплового стресса с большей вероятностью приобретают инсерции ONSEN, чем гены с повышенной экспрессией. Мы предполагаем, что тепловой стресс приводит к замене гистона H2A.Z гистоном H2A в термочувствительных генах (активируется во время теплового стресса), что снижает вероятность того, что ONSEN будет нацелен на эти гены. И наоборот, гены с пониженной регуляцией продолжают поддерживать гистон H2A.Z, который привлекает вставки ONSEN (Рисунок 16).

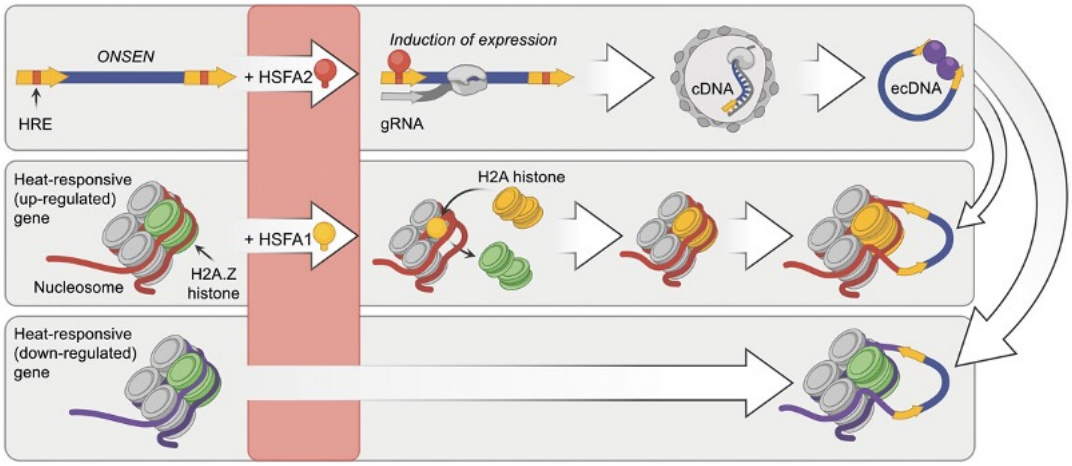

Рисунок 16. Модель, объясняющая более частые инсерции ONSEN в гены, активность которых снижается в ответ на тепловой стресс (37°C).

В целом, с помощью CANS мы смогли обнаружить изменения транспозиционной активности семейства транспозонов ONSEN, происходящие в ответ на тепловой стресс. Эта информация позволила нам определить «горячие точки» для инсерций ONSEN, а также показать, что с помощью CANS можно проводить детекцию унаследованных TEI. Это даёт возможность быстрой идентификации генов с инсерциями после активации МЭ в других растениях, включая агрономически важные культуры.

4.3.7 Мобильная активность экспрессирующихся ретротранспозонов подсолнечника

Чтобы проверить способность идентифицированных нами ранее exRTE подсолнечника генерировать новые инсерции (мобильная активность), было проведено сравнение относительного количества копий между сортами подсолнечника. Результаты показывают, что 25 (57%) exRTE имеют дополнительные копии в >1 сорте, что свидетельствует о потенциальной транспозиционной активности мобилома (группа 1). 19 (43%) exRTE имеют дополнительные копии в сортах 0–1 (группа 2; Рисунок 17A).

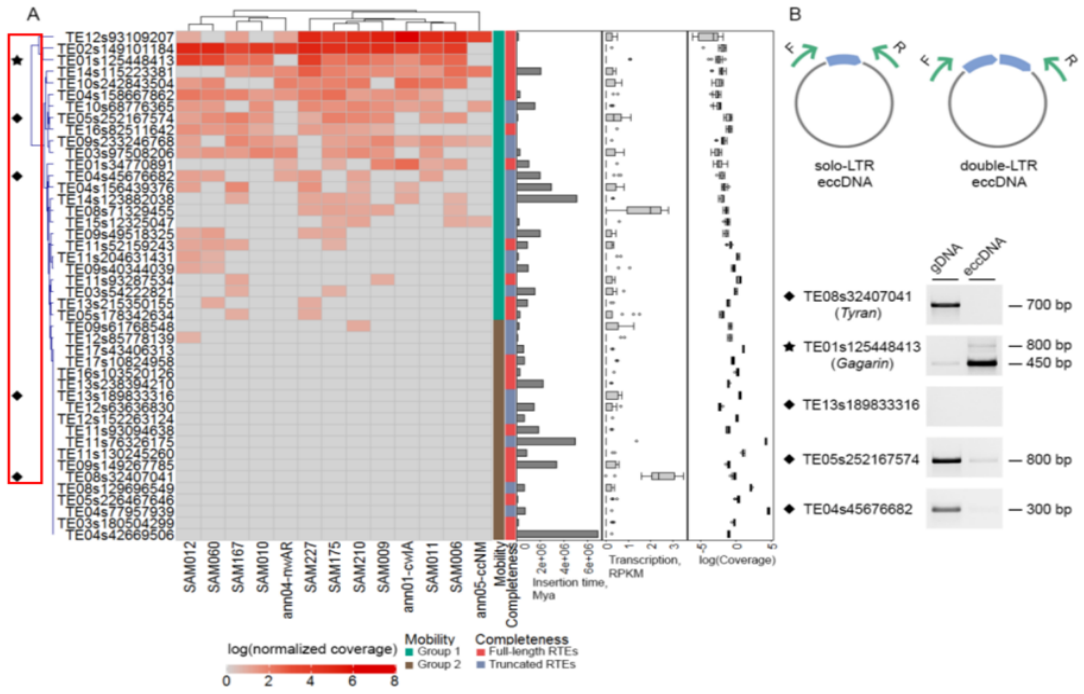

Рисунок 17. Мобильная активность exRTE. (A): Тепловая карта, демонстрирующая изменение соотношения между нормализованным покрытием RTE и минимальным покрытием ридами для этого RTE для 13 сортов подсолнечника (представлены коэффициент ACM, значения, преобразованные в log2). (B): инвертированная ПЦР с геномной ДНК (гДНК) и геномной ДНК, обогащенной внехромосомной кольцевыми ДНК (вкДНК). Звездочки указывают на RTE ‘Gagarin’.

Чтобы проверить текущую активность exRTE, был проведён анализ внехромосомных кольцевых ДНК (вкДНК) для пяти RTE, включая один с высокой и повсеместной экспрессией во всех проанализированных образцах, TE08s32407041HA412. Инвертированная ПЦР с использованием праймеров на exRTE TE01s125448413, названный нами «Gagarin», показала значительное увеличение концентрации ПЦР продукта в варианте с ДНК, обогащённой вкДНК (Рисунок 17B). Кроме того, были обнаружены вкДНК с одиночными и двойными LTR (Рисунок 17B). Примечательно, что RTE Gagarin является первым RTE подсолнечника с доказанной мобильной активностью.

Нами была также изучена мобильная активность exRTE, экспрессирующихся в ответ на эпигенетический стресс, как было показано нанопоровым секвенированием кДНК. Используя вышеописанный подход, мы определили, что 16 RTE (67%) имеют заметно увеличенное количество копий по крайней мере в одном сортообразце. ПЦР анализ вкДНК и инсерционного полиморфизма некоторых RTE позволили идентифицировать RTE с транспозиционной активностью под действием эпигенетического стресса. Это ретротранспозон *Ty3/Gypsy* был назван SUNTY3.

4.3.8 Новый *Ty1/Copia* LTR ретротранспозон MIG активен в развивающейся зерновке тритикале.

Анализ ретротранскриптома развивающейся зерновки тритикале выявил несколько экспрессирующихся LTR-RTE. Чтобы проверить их мобильную активность, мы определили образование внехромосомных кольцевых ДНК (вкДНК) этими RTE с помощью инвертированной ПЦР. Активность RTE3B нельзя было оценить с помощью инвертированной ПЦР. Мы продолжили детекцию вкДНК только для RTE7B. Для этого эксперимента была использована геномная ДНК, выделенная из развивающейся зерновки (10 dpa), а также колосовых и цветковых чешуй. Специфические продукты были обнаружены только для вкДНК, выделенной из развивающейся зерновки (Рисунок 18). Таким образом, наши результаты показали, что ретротранспозон RTE7B, названный нами "MIG", экспрессирует как изоформы РНК shGAG, так и изоформы геномной РНК и обладает транспозиционной активностью.

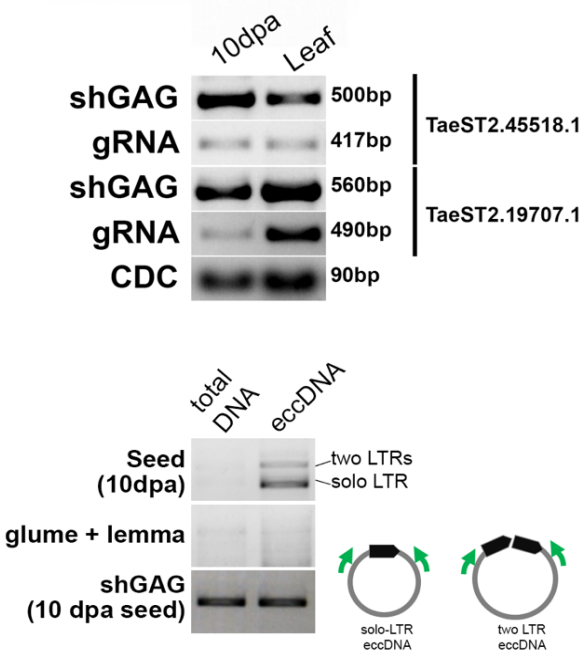

Рисунок 18. Экспрессия и формирование внехромосомной кольцевой ДНК (вкДНК). (A) Детекция с помощью ОТ-ПЦР изоформ shGAG и гРНК полноразмерных RTE RTE3B и RTE7B ('MIG'). CDC: референсный ген (белок, контролирующий клеточное деление). (B) Инвертированная ПЦР с фракцией, обогащенной геномной ДНК и вкДНК, полученной из развивающихся тканей семян (10 dpa) и колосковых и цветковых чешуй тритикале. Положения праймеров на вкДНК показаны на небольшом изображении справа. В качестве контроля использовали ПЦР с праймерами shGAG ДНК, обогащенной тотальной и вкДНК.

В заключение, в данной части работы было экспериментально доказано, что некоторые RTE-РНК происходят из автономных ретротранспозонов LTR с продолжающейся транспозиционной активностью на ранних стадиях развития зерновки тритикале. Полученные результаты закладывают фундамент для дальнейшего функционального изучения RTE-РНК и поиска полногеномных ассоциаций у тритикале и пшеницы.

4.3.9 Ретротранспозоны и эволюция размера генома *Fagopyrum tataricum* и *F. esculentum*

Род *Fagopyrum* включает примерно 25 видов, которые делятся на две клады: cymosum (крупная семянка) и urophyllum (маленькая семянка). Размеры генома двух ветвей значительно различаются, несмотря на одинаковое число хромосом (восемь) у большинства видов. Причины различий в размере генома остаются неизвестными. *F. esculentum* относится к крупносеменной кладе; его геном примерно в 3 раза больше (1,39 пг/1C), чем у близкородственного вида *F. tataricum* (0,56 пг/1C), а число хромосом у этих видов одинаковое. Полиплоидия и МЭ являются основными факторами изменчивости размера генома растений у разных видов. Поскольку оба вида являются диплоидами и имеют одинаковое число хромосом (2n = 2x = 16), было предположено, что активность МЭ может быть одной из причин увеличения размера генома *F. esculentum*. Чтобы получить предварительное представление о полногеномных различиях в составе повторов между двумя видами, была проведена сравнительная кластеризация геномных ридов Illumina с последующей аннотацией кластеров с использованием программного обеспечения RepeatExplorer. Характеристика повторов с помощью RepeatExplorer основана на геномных ридах и не зависит от качества сборки

генома, которое может немного различаться для двух видов. Этот анализ выявил схожее содержание в геноме сателлитных повторов (3,45% и 4%) и повторов класса II (ДНК-транспозоны, 1,97% и 1,98%), в то время как доля генома, занятая ретротранспозонами (RTE, повторы класса I), составила ~2 раз больше у *F. esculentum* (64,8%), чем у *F. tataricum* (30,7%) (Рисунок 19).

Затем мы воспользовались актуальной сборкой генома *F. esculentum*, чтобы получить более детальное представление о разнообразии LTR RTE, и сравнили эти данные с данными *F. tataricum*. Классификация RTE показала, что около 70% всех RTE у обоих видов принадлежат суперсемейству *Ty3/Gypsy*. Дальнейшая классификация всех RTE выявила поразительные различия в количестве копий отдельных семейств *Ty3/Gypsy* и *Ty1/Copia* между двумя видами (Рисунок 19B, C).

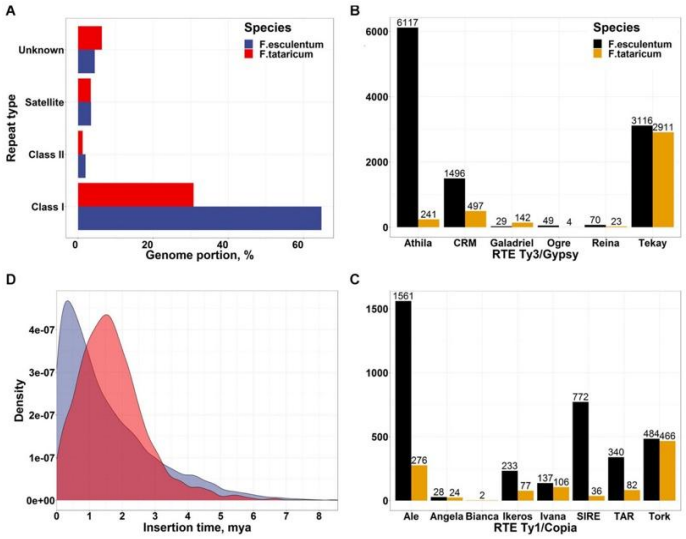

Рисунок 19. (A) Сравнение представленности в геноме *F. esculentum* и *F. tataricum* различных типов повторов на основе кластеризации геномных ридов с помощью RepeatExplorer. (B, C) Количество полноразмерных RTE *Ty3/Gypsy* и *Ty1/Copia* разных семейств на основе предсказания в собранных последовательностях генома. (D) Сравнение времени инсерции для *F. esculentum* (синий) и *F. tataricum* (красный) RTE.

Анализ семейств RTE в геноме показал, что *Athila Ty3/Gypsy* было наиболее распространенным семейством RTE (42% всех RTE) в геноме *F. esculentum*, в то время как *Tekay Ty3/Gypsy* было основным семейством (59% всех RTE) в *F. tataricum* геном. Сравнение копий RTE между геномами показало, что в геноме *F. esculentum* примерно в 25 раз больше элементов (6117 RTE) семейства *Athila Ty3/Gypsy*, чем у *F. tataricum* (241 RTE). Кроме того, семейства *Ale*, *SIRE*, *TAR* и *Ikeros Ty1/Copia* также показали значительные расхождения в количестве RTE между геномами двух видов. Это указывает на то, что RTE нескольких семейств накапливались с большей скоростью у *F. esculentum*, чем у *F. tataricum*. Затем мы провели анализ времени инсерции всех RTE у этих видов. Полученные оценки четко показали значительные различия (точный критерий Фишера для данных подсчета, p-значение < $2{,}2e^{-16}$) между видами, при этом 26,6% (3 834 полных RTE) RTE генома *F. esculentum* были интегрированы в геном во время последних 0,5 млн лет («недавние» вставки), в то время как только 9,6% RTE-вставок *F. tataricum* были классифицированы как «недавние» вставки.

Таким образом, характеристика повторяющихся элементов генома *F. esculentum*, показывает, что этот вид претерпел всплеск активности МЭ, произошедший < 0,5–1 млн лет назад, что и является одной из причин различий в размере генома с родственным видом *F. tataricum*.

4.4 Структура и состав внехромосомных кольцевых ДНК (вкДНК) LTR ретротранспозонов растений

ВкДНК представляют собой тип двухцепочечной ДНК, которая была обнаружена в клетках различных организмов, включая человека, животных и растения. Основным источником вкДНК в геномах растений являются МЭ. Появление высокопроизводительного секвенирования и биоинформатических методов способствовало пониманию разнообразия кольцевых молекул, составляющих так называемый циркулом растения. Было высказано предположение, что вкДНК МЭ генерируются путем гомологичной рекомбинации и негомологичного соединения концов линейной кДНК LTR ретротранспозонов. Исходя из этого, было предложено использовать вкДНК в качестве маркера мобильности МЭ. Ранее было проведено секвенирование вкДНК с помощью коротких ридов, что ограничивало изучение структуры и состава вкДНК. Секвенирование вкДНК длинными ридами широко не использовалось на растениях, оставляя неисследованными структурные особенности вкДНК растений. Целью этой части диссертационного исследования было изучить структуру и состав вкДНК у *A. thaliana* и *Brassica napus* с помощью нанопорового секвенирования вкДНК и использовать полученные сведения для идентификации новых активных транспозонов в геноме этих видов.

4.4.1 Особенности вкДНК *A. thaliana*, выявленные с помощью нанопорового секвенирования

Для проверки метода нанопорового секвенирования в качестве метода расшифровки вкДНК был использован мутант *ddm1 A. thaliana*, который обладает высоким уровнем транспозонной активности и хорошо изученным составом циркулома. Для обогащения вкДНК в образце ДНК использовали удаление линейной ДНК с помощью экзонуклеазы PlasmidSafe с последующей RCA амплификацией. Затем продукты RCA расщепляли с помощью эндонуклеазы T7 для их разделения (Рисунок 20A).

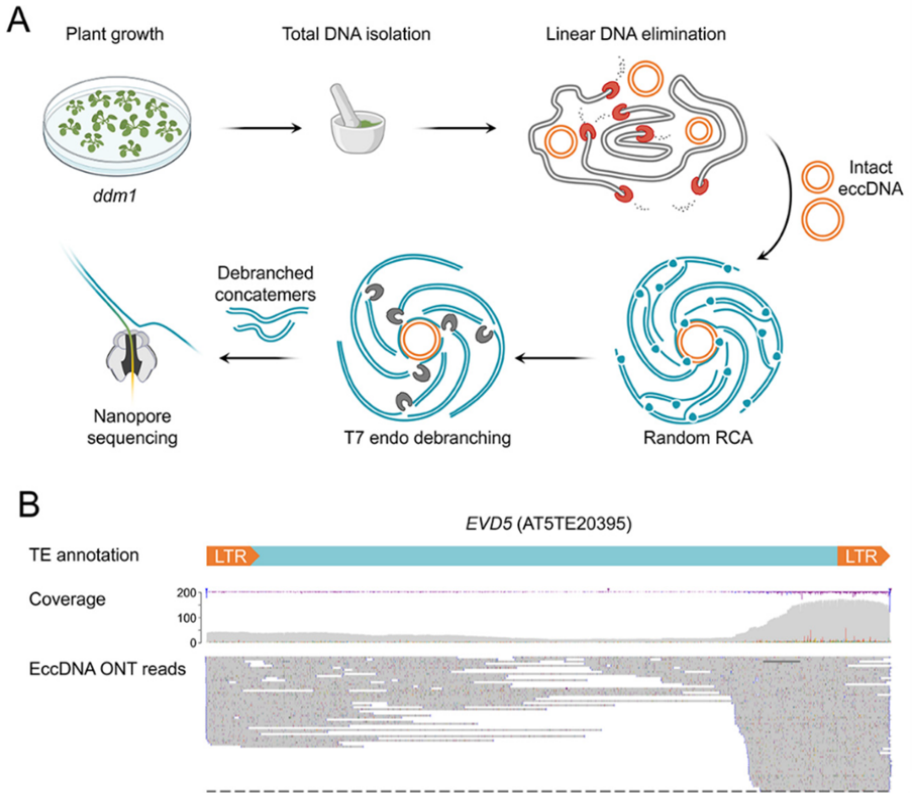

Рисунок 20. Нанопоровое секвенирование вкДНК мутанта *ddm1 A. thaliana*. (A) Схема эксперимента. (B) Покрытие EVD5 (AT5TE20395) ридами вкДНК ddm1.

Секвенирование вкДНК *ddm1* с помощью MinION позволило получить 257563 ридов. Мы использовали ONT-риды, содержащие две или более тандемно организованных мономерных единицы (последовательность одного круга молекулы вкДНК) длиной > 500 п.н., чтобы отфильтровать риды, которые не соответствовали продуктам RCA вкДНК. После этого скрининга 45277 (17,6%) ридов ONT остались для дополнительного изучения. Мы оценили соотношение log2 между количеством ридов для образца вкДНК *ddm1* относительно WGS *ddm1*, чтобы идентифицировать локусы с высоким покрытием ридами вкДНК. Были обнаружены четыре пика вкДНК с отношениями log2 > 3; два из них, SRC2 (AT1G09070.1) и MKK9 (AT1G73500), были пиками в дистальных областях хромосомы 1. Три копии LTR ретротранспозонов EVD1 (AT1TE41575), EVD2 (AT1TE41580) и EVD5 (AT5TE20395) из семейства ATCOPIA93 составили два других пика (Рисунок 21). Таким образом, мы показали, что нанопоровое секвенирование вкДНК является эффективным методом обнаружения локусов, продуцируемых вкДНК, таких как TEs и гены. Наши результаты показывают, что EVD5 является наиболее активным TE, продуцирующим вкДНК, в генотипе *ddm1*, что коррелирует с предыдущими исследованиями.

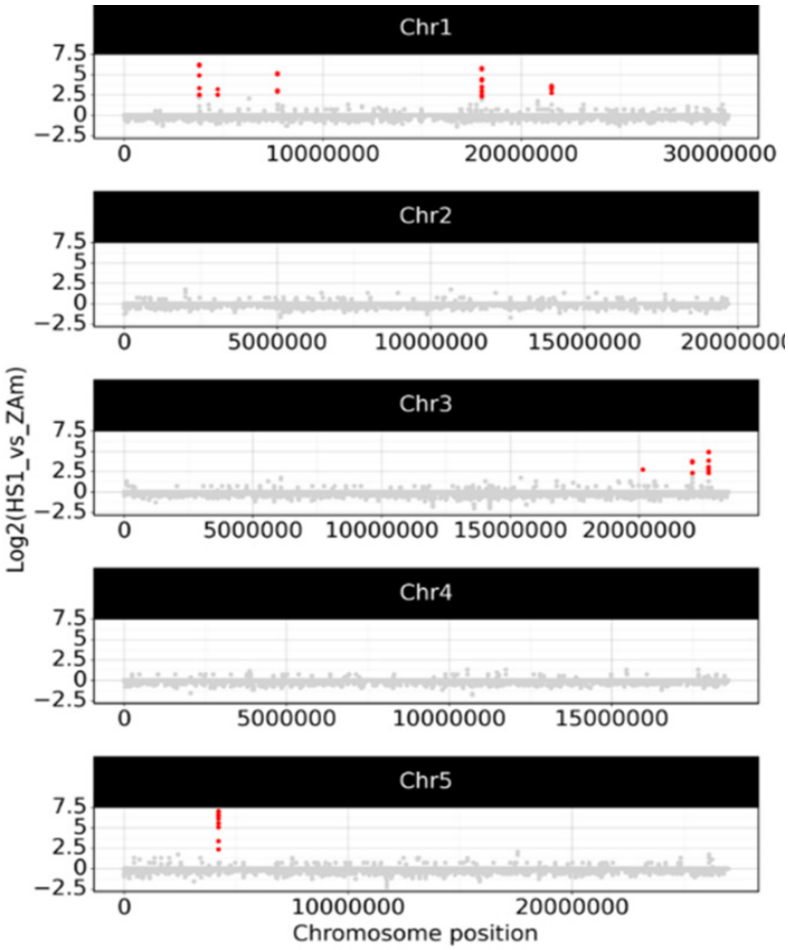

Рисунок 21. Локусы вкДНК, которые были чувствительны к тепловому стрессу. (А) покрытие геномных локусов ридами вкДНК в образце HS. Точки представляют отношение log2 числа картированных ридов вкДНК с конкатемером для образцов HS ZA и ZA (использовались только первичные выравнивания). В образцах HS ZA по сравнению с образцами ZA красными точками обозначены области генома со значительно увеличенным покрытием ридами (точный критерий Фишера с поправкой Бенджамини–Хохберга).

Предыдущие исследования показали, что у растений, выращенных на среде с добавлением зебуларина и аманитина (ZA) и подвергнутых стрессу, такому как тепловой, образуются вкДНК. Мы объединили ZA с каждым из следующих стимулов для выявления чувствительных к стрессу локусов вкДНК: тепловой стресс (HS), обработка флагеллином (Flg) и абсцизовой кислотой (ABA). Этот анализ не выявил никаких Flg- или ABA-чувствительных пиков вкДНК. После этого мы изучили данные ONT для образцов HS и обнаружили семь локусов, продуцирующих вкДНК, которые имели значительно более высокое покрытие конкатемерными ридами в образцах HS ZA, чем в образцах ZA (Рисунок 21). Эти пики соответствовали семи копиям LTR-ретротранспозона ONSEN семейства ATCOPIA78. Полученные данные свидетельствуют, что состав вкДНК слабо отличался от контроля для образцов ZA, ZA + ABA и ZA + Flg, тогда как применение ZA в комбинации с тепловым стрессом приводило к образованию вкДНК для семейства LTR-ретротранспозонов ONSEN.

Данные нанопорового секвенирования могут помочь установить полноразмерную последовательность вкДНК. Поэтому мы создали оригинальный конвейер (eccStructONT), позволяющий реконструировать последовательности вкДНК. Для элементов EVD и ONSEN (Рисунок 22) распределение длин вкДНК показало два пика, соответствующих полноразмерным (fl_вкДНК, около 5000 п.н.) и укороченным (tr_вкДНК, 1000 п.н.) молекулам вкДНК. Затем мы классифицировали различные молекулы вкДНК в зависимости от участка МЭ. Большинство вкДНК EVD были fl_вкДНК.

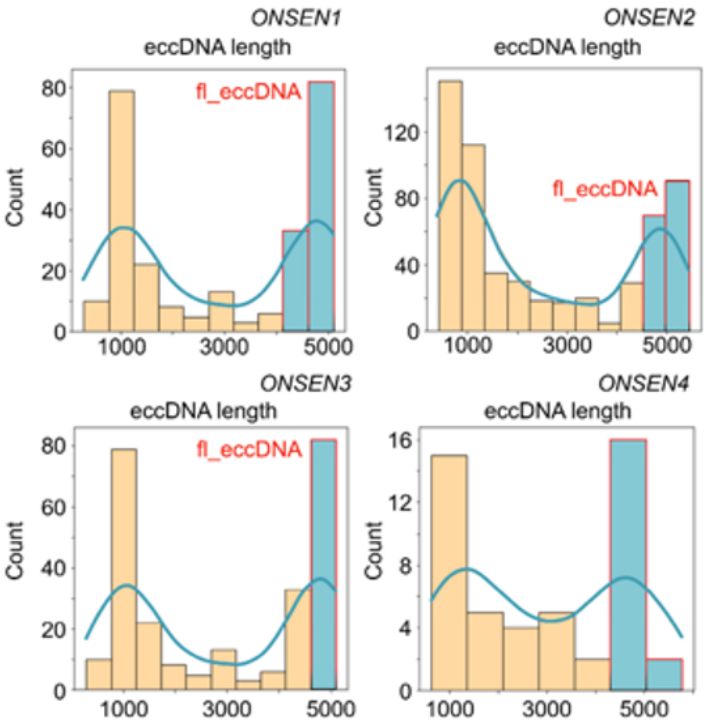

Рисунок 22. Структура вкДНК ONSEN, генерируемых в растениях *ddm1* в условиях умеренного теплового стресса. Гистограммы, показывающие количество молекул ONSEN вкДНК разной длины.

Визуализация отдельных структурных групп вкДНК для ONSEN позволяет предположить, что fl_вкДНК и tr_вкДНК были двумя основными формами вкДНК (Рисунок 22). ONSEN tr_вкДНК в основном происходят из одного или двух LTR. Соответственно, элементы EVD и ONSEN значительно различались по составу вкДНК, при этом ONSEN имел значительно большую долю tr_вкДНК, полученных из LTR, чем EVD.

Мы исследовали, остается ли соотношение fl_вкДНК к tr_вкДНК ONSEN таким же у растений дикого типа в условиях стресса ZA HS, как и у растений *ddm1* HS. Результаты показали, что что все шесть ONSEN образуют в основном tr_вкДНК с очень небольшим количеством fl_вкДНК. Опять же, LTR регионы ONSEN были основными источниками tr_вкДНК.

Полученные данные свидетельствуют, что состав вкДНК одного МЭ может значительно различаться в зависимости от эпигенетических факторов (мутация гена *DDM1*).

4.4.2 Нанопоровое секвенирование вкДНК рапса (*Brassica napus*) выявило новое семейство активных LTR ретротранспозонов ANTARES.

Рапс (*Brassica napus* L.) – природный амфидиплоид, образовавшийся в результате гибридизации диплоидных видов-предшественников *B. rapa* и *B. oleracea*. Рапс – вторая по важности масличная культура в мире после подсолнечника. Геном рапса секвенирован, и теперь доступно несколько сборок. Эти исследования показали, что 35 – 58,2% генома рапса занимают МЭ. Недавние исследования показали состав активных транспозонов рапса и их связь с важными показателями урожайности. Однако, на сегодня не известно ни одного МЭ рапса, активного в наши дни и в определенных условиях. В этой части диссертационной работы, мы стремились идентифицировать потенциально активные МЭ рапса и стимулировать их транспозицию, при этом обеспечивая высокую выживаемость растений.

Первоначально мы использовали оригинальный протокол активации транспозонов на основе использования токсинов AZ, но столкнулись с рядом трудностей. Во-первых, токсины подавляют развитие вторичных корней и апикальных меристем. Во-вторых, прямой посев семян на питательную среду, содержащую AZ в стандартных концентрациях (4 мкг/мл альфа-аманитина и 8 мкг/мл зебуларина (4+8 AZ)) приводил к гибели 95% проростков. Аналогичная проблема наблюдалась при применении технологии TE-генезиса на подсолнечнике, где не удалось получить жизнеспособные растения. В нашем опыте в таких условиях смогло вырасти только одно растение, но оно отличалось задержкой роста. Поэтому было решено адаптировать метод для рапса.

Мы разработали модификацию протокола TE-генезис. Согласно данной модификации стерильные семена сначала культивировали на среде Мурасиге-Скуга (MS), не содержащей токсинов. В этих условиях растения развивались в течение 7 дней до появления боковых корней. Затем их перенесли в среду MS, содержащую ингибиторы сайленсинга в стандартной концентрации (4+8 AZ) или сниженной наполовину для каждого ингибитора (2+4 AZ). Затем растения помещали в условия холодового стресса при +4°C (CS) в темноте на 1 день, после чего следовал 24-часовой тепловой стресс при +37°C (HS) перед возвращением в нормальные условия. Затем сеянцы адаптировали к условиям *ex vivo*. В результате по адаптированному протоколу были получены 6 (50%) и 8 (75%) растений для двух концентраций ((4+8 AZ) и (2+4 AZ), соответственно), что достаточно для анализа вкДНК. Эти растения отличались от контроля задержкой цветения и плодоношения на 14 дней. Однако все растения были полностью фертильны, что позволило получить популяцию M1.

Далее мы провели детекцию вкДНК рапса с помощью ONT секвенирования (Рисунок 23). Для обогащения вкДНК и последующего секвенирования были взяты следующие образцы: контрольные растения без AZ и стресса (NS_control), растения, выращенные на среде с AZ и без стресса (AZ_NS), растения с обработкой AZ и холодовым стрессом (4°C, 24 ч.) (AZ_CS) и растения с обработкой AZ, холодовым стрессом (4°C, 24 ч) и тепловым стрессом (37°C, 24 ч) (AZ_HS). Было обнаружено, что распределение ридов кольцевых ДНК по образцам не было равномерным: варианты AZ_NS и AZ_CS_HS показали самый высокий процент конкатемерных ридов общего пула - 32,26% и 20,01%, соответственно, тогда как варианты NS_control и AZ_CS показали 6,80% и 1,50%, соответственно.

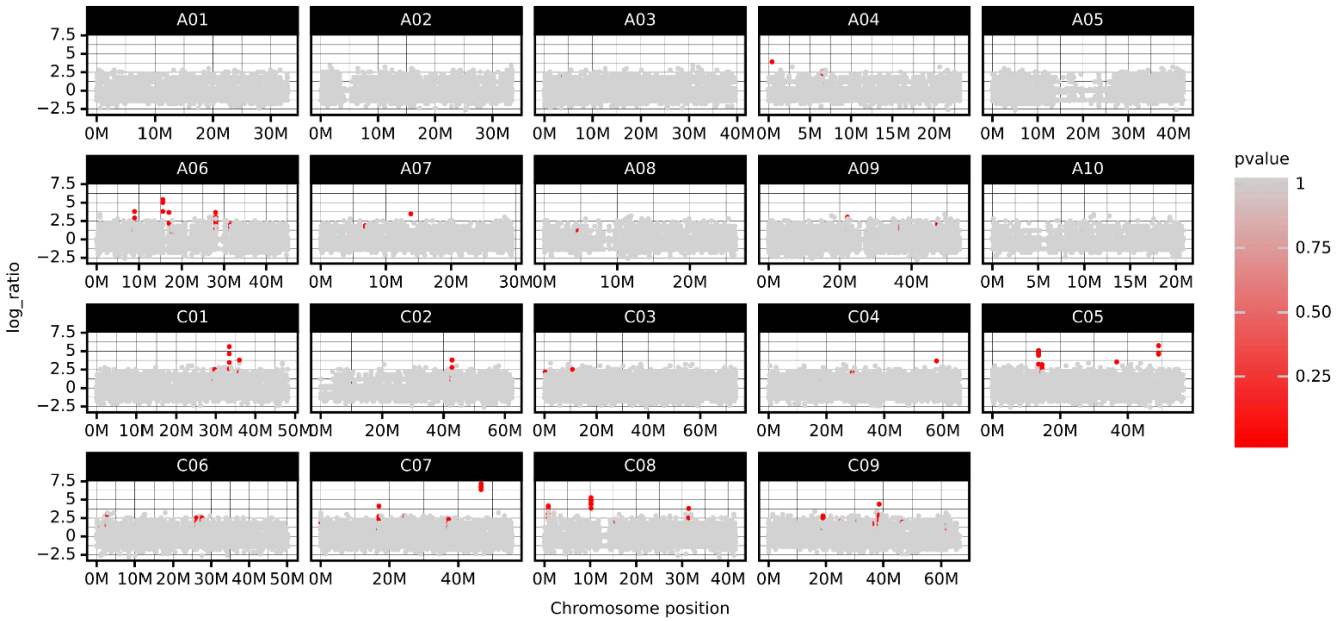

Рисунок 23. Покрытие геномных локусов ридами вкДНК в образце AZ_HS (среда с добавлением AZ и тепловой стресс). Точки представляют собой отношение log2 количества картированных чтений конкатемеров вкДНК для образцов AZ_HS и NS_control (использовались только первичные выравнивания), красные точки обозначают геномные области со значительно увеличенным покрытием ридами в образце AZ_HS (точный критерий Фишера с поправкой Бенджамини-Хохберга).

Дальнейший анализ выявил 6 LTR ретротранспозонов, продуцирующих вкДНК в ответ на обработку AZ и тепловой стресс (Рисунок 24A). Мы провели ПЦР с праймерами, специально разработанными для амплификации образцов вкДНК LTR-LTR. ПЦР с продуктом RCA и праймерами на идентифицированные МЭ привела к получению продуктов ПЦР, подтверждающих транспозиционную активность этих МЭ. Было обнаружено, что все локусы вкДНК локализованы в субгеноме C и являются представителями суперсемейства *Ty3/Gypsy*, клады *Tekay*. Представленные мобильные элементы активируются в ответ на тепловой стресс в сочетании с AZ-обработкой (AZ_HS) или только в ответ на AZ-обработку без стресса (NS_AZ). Варианты без стресса (NS_control) или обработки AZ и холодового стресса (AZ_CS) не показали значительного увеличения продукции вкДНК.

Структурный анализ вкДНК выявил гетерогенную по структуре популяцию молекул вкДНК в клетках рапса. На ряду с вкДНК, соответствующих полноразмерным TE ANTARES, были также детектированы вкДНК без одного LTR и сиквенсов, кодирующих один или все домены (Рисунок 24B).

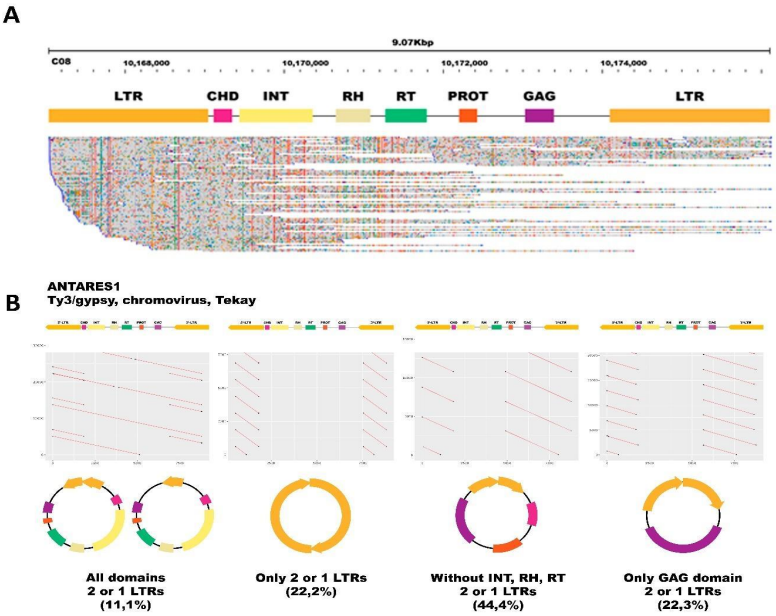

Рисунок 24. A – Покрытие локуса C08:10,167,043..10,176,109 ридами вкДНК AZ_HS; B - Схематическое изображение структуры вкДНК AZ_HS локуса C07: 46638253..46645089, показывающее части вкДНК: LTR (оранжевый), аспарагиновая протеаза (AP, красный), интеграза (INT, желтый), обратная транскриптаза (RT, зеленый) и РНКаза H (RH, фиолетовый), хромодомен (CHD, розовый).

Все геномные последовательности МЭ содержали родственные домены (Рисунок 24), необходимые для жизненного цикла мобильных элементов, за одним исключением: белок GAG отсутствовал в локусе C07: 46638253..46 645 089. Локус C08:815724..838250 характеризовался наличием трех тандемно организованных МЭ. Консенсусная последовательность LTR составляла 2142 п.н., а идентичность последовательности составляла 82,53%, что позволяло отнести эти элементы к одному семейству, названному нами ANTARES в честь сорта, в котором они были обнаружены.

4.5 Мобильные элементы и протеом растений

Активность МЭ сопряжена с трансляцией их белков и их взаимодействием между собой. Однако белки МЭ никто не изучал до этого с помощью прямой детекции методами масс-спектрометрии. Стоит предположить, что количество белков МЭ очень невелико в клетке в не стрессовых условиях из-за активности многочисленных систем сайленсинга. Поэтому для детекции белков МЭ необходимо иметь растения с повышенной активностью МЭ в результате нарушения их сайленсинга. Чтобы провести анализ белков МЭ мы использовали мутант *A. thaliana ddm1*. Для создания базы белковых последовательностей нами были использованы уникальные данные, которые были получены в результате прямого секвенирования РНК арабидопсиса. Это дало возможность точно предсказать ОРС для белков, принимая во внимание экзон-интронную структуру транспозонов и границы транскрипции. Используя собранный транскриптом, мы провели предсказание ОРС, включая короткие рамки считывания (кОРС, <300 пар нуклеотидов) и соответствующих им белковых последовательностей. Эти данные, наряду с референсным протеомом арабидопсиса из базы данных Araport 11 и данными нанопорового секвенирования мутанта *ddm1/rdr6/polV* были объединены в общую базу данных. Анализ масс-спектрометрических данных был проведён в программе MaxQuant.

В результате аннотации полученных масс-спектров, мы детектировали 41 639 пептидов для 4 600 белковых групп, из которых 109 белковых групп были найдены только у *ddm1*, в двух образцах (Рисунок 25). Отбирали только те белковые группы, которые были представлены 2 и более пептидами.

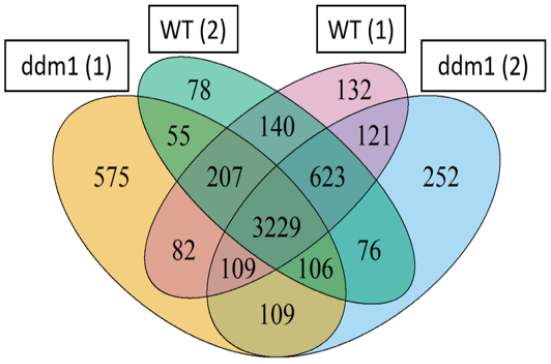

Рисунок 25. Диаграмма отношения числа белковых групп, детектированных в результате аннотации масс-спектров в программа MaxQuant.

Далее мы провели отбор белков, которые были выявлены только у *ddm1*. Среди них оказались как уже известные белки МЭ, принадлежащие ДНК транспозонам, так и ранее не описанные, трансляция которых идет с транскриптов транспозонов. Один из таких белков принадлежит транспозонам группы *VANDAL21*. По данным масс-спектрометрического анализ удалось детектировать только белок VanB у *ddm1*, несмотря на наличие транскриптов для остальных белков этого транспозона. Также было детектировано несколько пептидов VanB с других копий транспозона на хромосомах 4 и 5. Также нам удалось детектировать белки транспозазы различных копий транспозона *En/Spm*.

Масс-спектрометрический анализ позволил выявить и белки ретротранспозонов. Так, были детектированы белки ретротранспозонов семейства *Athila*. Интересно отметить, что детектированные белки транслировались с дополнительной ОРС и показывали сходство с белком вирусной частицы *env*, согласно данным BLASTp GyDB.

Зачастую при предсказании рамок считывания, минимальный размер предсказанной рамки начинается от 300 и более нуклеотидов. В результате чего, короткие рамки считывания менее 100 аминокислот не попадают в анализ. Однако большое число масс-спектрометрических исследований доказывает, что кОРС пептиды выполняют разные клеточные функции. Функции этих микропептидов остаются малоизученными, предположительно, часть из них выполняет регуляторную функцию. Из-за возможной функциональной значимости таких кОРС, нами также был проведён анализ их белков. При анализе нами были найдены пептиды с двух кОРС, которые, транслируются с дополнительной третьей рамки считывания ретротранспозона семейства *Ty3/Gypsy ATLANTYS1* с «минус» цепи. *ATLANTYS1* активируется в репродуктивных органах *A. thaliana* в условиях деметилирования.

Таким образом, проведённые исследования показывают, что мобильные элементы могут кодировать белки разной природы, включая белки с известной функцией (например, транспозаза и белок *Env*), а также неизвестные белки. К

последним относятся и продукт кОРС МЭ, функции которых остаются неизвестными.

5. ЗАКЛЮЧЕНИЕ

Проведённые в рамках данной работы биоинформатические и экспериментальные исследования разных видов растений (однодольные (*Allium cepa, A. fistulosum, x Triticosecale, Triticum aestivum*), двудольные (*Rosa wichurana, R. gallica, R. rugosa, R. foetida, R. chinensis, Helianthus annuus, Arabidopsis thaliana, Brassica napus*) и мхи (*Physcomitrium patens*)) показывают, что повторяющиеся элементы генома растений оказывают влияние на разные уровни организации, включая геном, циркулом, транскриптом и протеом клетки. Сами повторяющиеся элементы в свою очередь представляют уникальные инструменты для молекулярных и эволюционных исследований. В данной работе был создан арсенал методов для поиска и анализа сателлитных повторов (pyTanFinder, nanoTRF, DRAWID) и инсерций мобильных элементов (NanoCasTE, CANS, nanotei) в геномах растений. Используя созданные в данной работе и разработанные ранее методы был проведён широкомасштабный поиск новых сателлитных повторов в геномах однодольных (виды *Allium*), двудольных (виды *Rosa*) растений и мха (*P. patens*). Впервые идентифицированы новые высококопийные тандемные повторы для разных видов растений (*A. fistulosum*: AfiCen1K; *A. cepa*: TR2CL37, *R. wichurana* CL8, CL24; *R. chinensis*: CL226; 19 повторов для *P. patens*). Нами была изучена геномная организация идентифицированных сателлитных повторов и предложено их использование в качестве цитогенетических маркеров для хромосом соответствующих видов.

Благодаря адаптированию нанопорового секвенирования вкДНК и кДНК/РНК были также идентифицированы новые мобильные элементы с доказанной мобильной активностью в геноме тритикале (ретротранспозон 'MIG'), подсолнечника (ретротранспозоны 'Gagarin' и 'SUNTY3'), рапса (семейство ретротранспозонов 'Antares') и арабидопсиса ('TR-GAG' элемент). Эта часть работы представляет большой интерес для дальнейшего изучения эволюции и биологии идентифицированных сателлитных повторов и активных мобильных элементов.

В рамках транскриптомного анализа, в работе был проведен поиск доказательств экспрессии сателлитных повторов и LTR ретротранспозонов широкого круга растений на разных стадиях развития и под различными стрессовыми условиями. Эти исследования однозначно показывают, что оба типа повторов транскрибируются у растений, что было доказано как методами нанопорового секвенирования РНК/кДНК, так и RNAseq анализом и ОТ-ПЦР. В целом, полученные результаты показывают, что LTR ретротранспозоны, обладающие определёнными геномными характеристиками (удаленность от генов, размер рамок считывания, время инсерции в геном), вносят существенный вклад в транскриптом растений. У разных видов растений, включая однодольные и двудольные растения, наиболее представленными транскриптами LTR ретротранспозонов были транскрипты, кодирующие GAG белки. Эти транскрипты

транскрибируются как с полноразмерных, так и с неавтономных (TR-GAG) LTR ретротранспозонов, а также с "доместицированных" копий GAG генов.

Используя CANS и NanoCasTE, были идентифицированы особенности транспозиции в геноме *A. thaliana* двух ретротранспозонов, EVD и ONSEN. Показано, что инсерции EVD преимущественно образуются в прицентромерных областях в геноме мутантного растения *ddm1*, что отличается от экотипов *A. thaliana* дикого типа. Это указывает на ключевую роль метилирования ДНК в локализации инсерций EVD. Кроме того, проведённые исследования показывают, что инсерции ретротранспозонов семейства ONSEN, активированных в ответ на тепловой стресс, чаще происходят в гены, которые снижают свою экспрессию во время теплового стресса. Практически значимым результатом работы является показанная возможность детекции инсерций как МЭ, так и Т-ДНК с помощью CANS, что позволяет проводить быструю идентификацию генов с инсерциями у растений, включая сельскохозяйственные виды растений. В целом, CANS и NanoCasTE являются эффективными инструментами для идентификации мест инсерций МЭ у различных видов растений, что позволит в дальнейшем детально изучить активность мобилома в ответ на стресс.

Проведенные масс-спектрометрические исследования протеома растений *A. thaliana* показывают, что мобильные элементы могут кодировать белки разной природы, включая белки с известной функцией (например, транспозаза и белок *Env*), а также неизвестные белки, функции которых только предстоит понять в будущем.

Таким образом, в ходе проведённых исследований были созданы новые биоинформатические и молекулярные методы, которые позволяют проводить идентификацию и изучение повторяющихся элементов генома растений, используя современные геномные, цитогенетические, транскриптомные и циркуломные данные. Использование новых и уже существующих подходов совместно с современными омиксными данными позволило впервые идентифицировать новые активные мобильные элементы и сателлитные повторы разных видов растений, а также изучить их геномные, транскриптомные, протеомные и циркуломные (вкДНК) особенности. Идентифицированные активные мобильные элементы и, что важно, условия их активации в геноме закладывают основу для разработки технологии контролируемой активации мобилома для создания коллекций сельскохозяйственных растений с новыми инсерциями. Такие коллекции растений и новые методы детекции инсерций (включая CANS) будут ценным источником новых признаков для селекции растений и материалом для функциональной геномики.

6. ВЫВОДЫ

1. Новые высококопийные сателлитные повторы (*Allium fistulosum*: AfiCen1K; *Allium cepa*: TR2CL37, *Rosa wichurana* CL8, CL24; *Rosa chinensis*: CL226; 19 повторов для *Physcomitrium patens*) ассоциированы с перицентромерными, центромерными и гетерохроматиновыми регионами как у однодольных и двудольных растений, так и мха, и могут быть использованы в качестве хромосомных маркеров и для интегрирования геномных и цитогенетических карт.
2. Центромеры хромосом *Allium fistulosum* содержат инсерции хлоропластной ДНК и длинный (~1,25 т.п.н.) транскрибирующийся тандемный повтор, который отличается по хромосомной и геномной организации от центромерного повтора *A. cepa*.
3. Экспрессирующиеся LTR ретротранспозоны представлены в транскриптомах однодольных и двудольных видов растений на разных стадиях развития в нормальных и стрессовых условиях, а их геномная организация отличается от неэкспрессирующихся LTR ретротранспозонов, включая время инсерции, близость к генам, число копий и обогащение открытыми рамками считывания, кодирующими домен GAG.
4. Данные нанопорового секвенирования транскриптома, генома и внехромосомных кольцевых ДНК и их анализ с помощью разработанных программ (NanoCasTE, nanotei, eccStructONT) являются удобным инструментом для изучения мобилома растений, используя который, были впервые идентифицированы LTR ретротранспозоны тритикале (ретротранспозон 'MIG'), подсолнечника (ретротранспозоны 'Gagarin' и 'SUNTY3'), рапса (семейство ретротранспозонов 'Antares') и арабидопсиса ('TR-GAG' элемент), обладающие мобильной активностью в лабораторных условиях.
5. Инсерции LTR ретротранспозонов в геноме *Arabidopsis thaliana* возникают не в случайных локусах, а связаны с определёнными хромосомными (центромерные регионы) регионами, а также эпигенетическими и транскриптомными особенностями, как было показано с помощью разработанного CANS/NanoCasTE подхода и полногеномного анализа соматических инсерций EVD и ONSEN.
6. Мобильные элементы в ходе жизненного цикла образуют пул внехромосомных кольцевых ДНК, гетерогенных по структуре и композиции, а также кодируют набор белков с каноническими и пока неизвестными функциями.

Список публикаций по теме диссертации

Статьи в международных рецензируемых журналах

1. Merkulov, P.; Gvaramiya, S.; Dudnikov, M.; Komakhin, R.; Omarov, M.; Kocheshkova, A.; Konstantinov, Z.; Soloviev, A.; Karlov, G.; Divashuk, M.; **Kirov, I.** Cas9-Targeted Nanopore Sequencing Rapidly Elucidates the Transposition Preferences and DNA Methylation Profiles of Mobile Elements in Plants. *Journal of Integrative Plant Biology* **2023**, *65*, 2242–2261.
2. Polkhovskaya, E.; Bolotina, A.; Merkulov, P.; Dudnikov, M.; Soloviev, A.; **Kirov, I.** Long-Read cDNA Sequencing Revealed Novel Expressed Genes and Dynamic Transcriptome Landscape of *Triticale* (x *Triticosecale* Wittmack) Seed at Different Developing Stages. *Agronomy* **2023**, *13*, 292.
3. **Kirov, I.** Toward Transgene-Free Transposon-Mediated Biological Mutagenesis for Plant Breeding. *International Journal of Molecular Sciences* **2023**, *24*, 17054.
4. **Kirov, I.**; Kolganova, E.; Dudnikov, M.; Yurkevich, O.Y.; Amosova, A.V.; Muravenko, O.V. A Pipeline NanoTRF as a New Tool for De Novo Satellite DNA Identification in the Raw Nanopore Sequencing Reads of Plant Genomes. *Plants* **2022**, *11*, 2103.
5. **Kirov, I.**; Merkulov, P.; Polkhovskaya, E.; Konstantinov, Z.; Kazancev, M.; Saenko, K.; Polkhovskiy, A.; Dudnikov, M.; Garibyan, T.; Demurin, Y.; et al. Epigenetic Stress and Long-Read cDNA Sequencing of Sunflower (*Helianthus annuus* L.) Revealed the Origin of the Plant Retrotranscriptome. *Plants* **2022**, *11*, 3579.
6. Penin, A.A.; Kasianov, A.S.; Klepikova, A.V.; **Kirov, I.V.**; Gerasimov, E.S.; Fesenko, A.N.; Logacheva, M.D. High-Resolution Transcriptome Atlas and Improved Genome Assembly of Common Buckwheat, *Fagopyrum esculentum*. *Frontiers in plant science* **2021**, *12*, 612382.
7. **Kirov, I.**; Merkulov, P.; Dudnikov, M.; Polkhovskaya, E.; Komakhin, R.A.; Konstantinov, Z.; Gvaramiya, S.; Ermolaev, A.; Kudryavtseva, N.; Gilyok, M.; et al. Transposons Hidden in *Arabidopsis thaliana* Genome Assembly Gaps and Mobilization of Non-Autonomous LTR Retrotransposons Unravelled by Nanotei Pipeline. *Plants* **2021**, *10*, 2681.
8. **Kirov, I.**; Odintsov, S.; Omarov, M.; Gvaramiya, S.; Merkulov, P.; Dudnikov, M.; Ermolaev, A.; Laere, K.V.; Soloviev, A.; Khrustaleva, L. Functional *Allium fistulosum* Centromeres Comprise Arrays of a Long Satellite Repeat, Insertions of Retrotransposons and Chloroplast DNA. *Frontiers in Plant Science* **2020**, *11*, 562001.
9. **Kirov, I.**; Omarov, M.; Merkulov, P.; Dudnikov, M.; Gvaramiya, S.; Kolganova, E.; Komakhin, R.; Karlov, G.; Soloviev, A. Genomic and Transcriptomic Survey Provides New Insight into the Organization and Transposition Activity of Highly Expressed LTR Retrotransposons of Sunflower (*Helianthus annuus* L.). *International journal of molecular sciences* **2020**, *21*, 9331.
10. **Kirov, I.**; Dudnikov, M.; Merkulov, P.; Shingaliev, A.; Omarov, M.; Kolganova, E.; Sigaeva, A.; Karlov, G.; Soloviev, A. Nanopore RNA Sequencing Revealed Long Non-Coding and LTR Retrotransposon-Related RNAs Expressed at Early Stages of Triticale seed Development. *Plants* **2020**, *9*, 1794.
11. Fesenko, I.; **Kirov, I.**; Filippova, A. Impact of Noncoding Part of the Genome on the Proteome Plasticity of the Eukaryotic Cell. *Russian Journal of Bioorganic Chemistry* **2018**, *44*, 397–402.
12. Saint-Oyant, L.H.; Ruttink, T.; Hamama, L.; **Kirov, I.**; Lakhwani, D.; Zhou, N.-N.; Bourke, P.; Daccord, N.; Leus, L.; Schulz, D.; et al. A High-Quality Genome Sequence of *Rosa chinensis* to Elucidate Ornamental Traits. *Nature plants* **2018**, *4*, 473–484.
13. **Kirov, I.**; Gilyok, M.; Knyazev, A.; Fesenko, I. Pilot Satellitome Analysis of the Model Plant, *Physcomitrella patens*, Revealed a Transcribed and High-Copy IGS Related Tandem Repeat. *Comparative Cytogenetics* **2018**, *12*, 493.
14. Van Laere K.; Van Huylenbroeck J.; **Kirov I.**, Khrustaleva L. Cytogenetic approaches enhance advanced breeding in woody ornamental species. *Acta Horticulturae* **2018**, 1191, 9-16.
15. **Kirov, I.**; Khrustaleva, L.; Laere, K.V.; Soloviev, A.; Meeus, S.; Romanov, D.; Fesenko, I. DRAWID: User-Friendly Java Software for Chromosome Measurements and Idiogram Drawing. *Comparative cytogenetics* **2017**, *11*, 747.

16. **Kirov, I.V.**; Kiseleva, A.V.; Laere, K.V.; Roy, N.V.; Khrustaleva, L.I. Tandem Repeats of *Allium fistulosum* Associated with Major Chromosomal Landmarks. *Molecular Genetics and Genomics* **2017**, *292*, 453–464.
17. **Kirov, I.V.**; Laere, K.V.; Roy, N.V.; Khrustaleva, L.I. Towards a FISH-Based Karyotype of *Rosa* L. (Rosaceae). *Comparative Cytogenetics* **2016**, *10*, 543.
18. **Kirov I.V.**; Khrustaleva L.I.; Van Laere K.; Van Roy N. Molecular cytogenetics in the genus *Rosa*: current status and future perspectives. *Acta Horticulturae* **2015**, 1087, 41-48.

Патенты на изобретение

19. **Киров И.В.**, Меркулов П.Ю., Константинов З.С., Дудников М.В., Соловьев А.А., Карлов Г.И., Дивашук М.Г. (2022) Способ определения геномной локализации и числа копий Т-ДНК в геноме трансформированных растений с помощью Cas9-целевого нанопорового секвенирования. Патент РФ RU2785923.

Другие публикации в международных рецензируемых журналах

20. Merkulov, P.; Serganova, M.; Petrov, G.; Mityukov, V. and **Kirov, I.** Long-read sequencing of extrachromosomal circular DNA and genome assembly of a *Solanum lycopersicum* breeding line revealed active LTR retrotransposons originating from *S. peruvianum* L. introgressions. *BMC genomics* **2024** 25(1), 1-11.
21. Polkhovskaya, E.; Gruzdev, I.; Moskalev, E.; Merkulov, P.; Bolotina, A.; Soloviev, A.; **Kirov, I.** Nanopore Amplicon Sequencing Allows Rapid Identification of Glutenin Allelic Variants in a Wheat Collection. *Agronomy* **2023**, *14*, 13.
22. Sidorova, T.; Miroshnichenko, D.; **Kirov, I**.; Pushin, A.; Dolgov, S. Evaluation of the Eff Ect of the Transgenic Component of the Graft-Twin Combination on Resistance to the Plum Pox Virus. *Horticulture and viticulture* **2022**, 15–29.
23. Ermolaev, A.; Kudryavtseva, N.; Pivovarov, A.; **Kirov, I.**; Karlov, G.; Khrustaleva, L. Integrating Genetic and Chromosome Maps of *Allium* Cepa: From Markers Visualization to Genome Assembly Verification. *International journal of molecular sciences* **2022**, *23*, 10486.
24. Kudryavtseva, N.; Ermolaev, A.; Karlov, G.; **Kirov, I.**; Shigyo, M.; Sato, S.; Khrustaleva, L. A Dual-Color Tyr-FISH Method for Visualizing Genes/Markers on Plant Chromosomes to Create Integrated Genetic and Cytogenetic Maps. *International journal of molecular sciences* **2021**, *22*, 5860.
25. Sidorova, T.; Miroshnichenko, D.; **Kirov, I.**; Pushin, A.; Dolgov, S. Effect of Grafting on Viral Resistance of Non-Transgenic Plum Scion Combined with Transgenic PPV-Resistant Rootstock. *Frontiers in Plant Science* **2021**, *12*, 621954.
26. **Kirov, I.**; Polkhovskaya, E.; Dudnikov, M.; Merkulov, P.; Vlasova, A.; Karlov, G.; Soloviev, A. Searching for a Needle in a Haystack: Cas9-Targeted Nanopore Sequencing and DNA Methylation Profiling of Full-Length Glutenin Genes in a Big Cereal Genome. *Plants* **2021**, *11*, 5.
27. Bolsheva, N.L.; Melnikova, N.V.; **Kirov, I.V.**; Dmitriev, A.A.; Krasnov, G.S.; Amosova, A.V.; Samatadze, T.E.; Yurkevich, O.Y.; Zoshchuk, S.A.; Kudryavtseva, A.V.; et al. Characterization of Repeated DNA Sequences in Genomes of Blue-Flowered Flax. *BMC Evolutionary Biology* **2019**, *19*, 79–88.
28. Khrustaleva, L.; Kudryavtseva, N.; Romanov, D.; Ermolaev, A.; **Kirov, I.** Comparative Tyramide-FISH Mapping of the Genes Controlling Flavor and Bulb Color in *Allium* Species Revealed an Altered Gene Order. *Scientific reports* **2019**, *9*, 12007.
29. Fesenko, I.; **Kirov, I.**; Kniazev, A.; Khazigaleeva, R.; Lazarev, V.; Kharlampieva, D.; Grafskaia, E.; Zgoda, V.; Butenko, I.; Arapidi, G.; et al. Distinct Types of Short Open Reading Frames Are Translated in Plant Cells. *Genome research* **2019**, *29*, 1464–1477.

30. **Kirov, I.**; Pirsikov, A.; Milyukova, N.; Dudnikov, M.; Kolenkov, M.; Gruzdev, I.; Siksin, S.; Khrustaleva, L.; Karlov, G.; Soloviev, A. Analysis of Wheat Bread-Making Gene (Wbm) Evolution and Occurrence in Triticale Collection Reveal Origin via Interspecific Introgression into Chromosome 7AL. *Agronomy* **2019**, *9*, 854.
31. Fesenko, I.; Azarkina, R.; **Kirov, I.**; Kniazev, A.; Filippova, A.; Grafskaia, E.; Lazarev, V.; Zgoda, V.; Butenko, I.; Bukato, O.; et al. Phytohormone Treatment Induces Generation of Cryptic Peptides with Antimicrobial Activity in the Moss *Physcomitrella patens*. *BMC plant biology* **2019**, *19*, 1–16.
32. Kroupin, P.; Kuznetsova, V.; Romanov, D.; Kocheshkova, A.; Karlov, G.; Dang, T.X.; Khuat, T.M.L.; **Kirov, I.**; Alexandrov, O.; Polkhovskiy, A.; et al. Pipeline for the Rapid Development of Cytogenetic Markers Using Genomic Data of Related Species. *Genes* **2019**, *10*, 113.
33. Filippova, A.; Lyapina, I.; **Kirov, I.**; Zgoda, V.; Belogurov, A.; Kudriaeva, A.; Ivanov, V.; Fesenko, I. Salicylic Acid Influences the Protease Activity and Posttranslation Modifications of the Secreted Peptides in the Moss *Physcomitrella patens*. *Journal of Peptide Science* **2019**, *25*, e3138.
34. Fesenko, I.; Khazigaleeva, R.; **Kirov, I.**; Kniazev, A.; Glushenko, O.; Babalyan, K.; Arapidi, G.; Shashkova, T.; Butenko, I.; Zgoda, V.; et al. Alternative Splicing Shapes Transcriptome but Not Proteome Diversity in *Physcomitrella patens*. *Scientific reports* **2017**, *7*, 2698.
35. Bolsheva, N.L.; Melnikova, N.V.; **Kirov, I.V.**; Speranskaya, A.S.; Krinitsina, A.A.; Dmitriev, A.A.; Belenikin, M.S.; Krasnov, G.S.; Lakunina, V.A.; Snezhkina, A.V.; et al. Evolution of Blue-Flowered Species of Genus *Linum* Based on High-Throughput Sequencing of Ribosomal RNA Genes. *BMC evolutionary biology* **2017**, *17*, 23–36.
36. Yurkevich, O.Y.; **Kirov, I.V.**; Bolsheva, N.L.; Rachinskaya, O.A.; Grushetskaya, Z.E.; Zoschuk, S.A.; Samatadze, T.E.; Bogdanova, M.V.; Lemesh, V.A.; Amosova, A.V.; et al. Integration of Physical, Genetic, and Cytogenetic Mapping Data for Cellulose Synthase (CesA) Genes in Flax (*Linum usitatissimum* L.). *Frontiers in Plant Science* **2017**, *8*, 288475.
37. Laskowska, D.; Berbeć, A.; Laere, K.V.; **Kirov, I.**; Czubacka, A.; Trojak-Goluch, A. Cytology and Fertility of Amphidiploid Hybrids between *Nicotiana* Wuttkei Clarkson et Symon and *N. tabacum* L. *Euphytica* **2015**, *206*, 597–608.
38. Divashuk, M.G.; Khuat, T.M.L.; Kroupin, P.Y.; **Kirov, I.V.**; Romanov, D.V.; Kiseleva, A.V.; Khrustaleva, L.I.; Alexeev, D.G.; Zelenin, A.S.; Klimushina, M.V.; et al. Variation in Copy Number of *Ty3/Gypsy* Centromeric Retrotransposons in the Genomes of *Thinopyrum intermedium* and Its Diploid Progenitors. *PLoS One* **2016**, *11*, e0154241.
39. **Kirov, I.V.**; Laere, K.V.; Khrustaleva, L.I. High Resolution Physical Mapping of Single Gene Fragments on Pachytene Chromosome 4 and 7 of *Rosa*. *BMC genetics* **2015**, *16*, 1–10.
40. **Kirov, I.**; Divashuk, M.; Laere, K.V.; Soloviev, A.; Khrustaleva, L. An Easy "SteamDrop" Method for High Quality Plant Chromosome Preparation. *Molecular cytogenetics* **2014**, *7*, 1–10.
41. **Kirov, I.**; Laere, K.V.; Riek, J.D.; Keyser, E.D.; Roy, N.V.; Khrustaleva, L. Anchoring Linkage Groups of the *Rosa* Genetic Map to Physical Chromosomes with Tyramide-FISH and EST-SNP Markers. *PloS one* **2014**, *9*, e95793.
42. Kiseleva, A.; **Kirov, I.**; Khrustaleva, L. Chromosomal Organization of Centromeric *Ty3/Gypsy* Retrotransposons in *Allium cepa* L. and *Allium* Fistulosum L. *Russian journal of genetics* **2014**, *50*, 586–592.
43. Divashuk, M.G.; Alexandrov, O.S.; Razumova, O.V.; **Kirov, I.V.**; Karlov, G.I. Molecular Cytogenetic Characterization of the Dioecious *Cannabis sativa* with an XY Chromosome Sex Determination System. *PloS one* **2014**, *9*, e85118.