На правах рукописи

БИРЮКОВ МИХАИЛ ЮРЬЕВИЧ

ПРОИСХОЖДЕНИЕ И ЭВОЛЮЦИЯ СТРУКТУРНЫХ ВАРИАНТОВ *Tat* LTR-РЕТРОТРАНСПОЗОНОВ ЗЕЛЁНЫХ РАСТЕНИЙ

1.5.7. – генетика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата биологических наук

Новосибирск 2023

Работа выполнена в межинститутской лаборатории молекулярной палеогенетики и палеогеномики развития Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», г. Новосибирск.

Научный руководитель:	**Устьянцев Кирилл Валерьевич**, кандидат биологических наук, научный сотрудник сектора молекулярно-генетических механизмов регенерации ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики СО РАН», г. Новосибирск.
Официальные оппоненты:	**Нетёсов Сергей Викторович**, доктор биологических наук, Заведующий Лабораторией бионанотехнологии, микробиологии и вирусологии ФЕН НГУ, г. Новосибирск
	Щербаков Дмитрий Юрьевич, доктор биологических наук, Заведующий лабораторией геносистематики ФГБУН «Лимнологический институт СО РАН», г. Иркутск
Ведущая организация:	ФГБУН «Институт молекулярной и клеточной биологии СО РАН», г. Новосибирск

Защита состоится: “__” _____ 2023 г. на утреннем заседании диссертационного совета 24.1.239.01 на базе ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики СО РАН» в конференц-зале Института по адресу: пр. академика Лаврентьева 10, г. Новосибирск, 630090, тел: +7(383) 363-49-06 (1321); email: dissov@bionet.nsc.ru, факс: +7(383) 333-12-78

С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте Института: www.icgbio.ru

Автореферат разослан “__” _____ 2023 г.

Учёный секретарь

диссертационного совета,

доктор биологических наук Хлебодарова Т.М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность.

Мобильные генетические элементы (МГЭ) - последовательности генома, способные менять своё положение в нём. Благодаря этому свойству МГЭ являются важными регуляторами структуры и функционирования генома. Среди МГЭ выделяют широкую группу - класс ретротранспозоны. Отличительной особенностью ретротранспозонов является механизм перемещения по принципу “копирование и вставка” (“copy-and-paste”), в основе которого лежит процесс обратной (“ретро-”) транскрипции. В процессе обратной транскрипции по матрице мРНК ретротранспозона синтезируется его новая ДНК-копия, которая затем встраивается в новое место в геноме. Благодаря этому свойству ретротранспозоны могут составлять подавляющую долю от размеров геномов эукариот, способную достигать 90% в случае геномов некоторых высших растений.

Среди ретротранспозонов выделяют отдельный большой подкласс элементов с длинными концевыми повторами (LTR - long terminal repeat) - LTR-ретротранспозоны. LTR-ретротранспозоны структурно и эволюционно близки ретровирусам позвоночных животных, а также обладают схожим жизненным циклом. Основная разница состоит лишь в том, что ретровирусы способны покидать клетку организма-хозяина и заражать новые клетки, в то время как жизненный цикл LTR-ретротранспозонов полностью ограничен клеткой-хозяином.

Ранее в нашей лаборатории были исследованы элементы кластера *Tat* группы LTR-ретротранспозонов Ty3/gypsy, распространенного исключительно в геномах зелёных растений (Viridiplantae). Отличительной особенностью *Tat* LTR-ретротранспозонов от других LTR-элементов является наличие дополнительного домена рибонуклеазы H (RNH) в гене полипротеина (pol). Было показано, что у элементов в данном кластере существуют различные структурные варианты, отличающиеся как положением дополнительного домена RNH внутри гена pol, так и присутствием/отсутствием дополнительных открытых рамок считывания с неясной функцией. Причём, разнообразие структур *Tat* LTR-ретротранспозонов коррелирует с вертикальной эволюцией крупных таксонов зелёных растений. Однако все эти данные были получены преимущественно на геномах наиболее высокоорганизованных групп цветковых растений, а разнообразие таксонов нецветковых растений ограничивалось лишь одним геномом плауна (Selaginella moellendorffii; класс Lycopodiopsida) и тремя геномами хвойных (класс Pinopsida, семейство Pinaceae).

К настоящему моменту значительно возросло количество геномных сборок представителей таксонов нецветковых и древних таксонов цветковых растений,

открывая возможность для более детального исследования эволюции *Tat* LTR-ретротранспозонов на ранних этапах их дивергенции. Кроме того, это может пролить свет на происхождение и распространение дополнительного домена RNH у данных элементов.

Целью данной работы является изучение происхождения и эволюции структурных вариантов *Tat* LTR-ретротранспозонов зелёных растений.

Задачи:

1. Разработать методологические подходы к углублённому изучению LTR-ретротранспозонов.
2. Исследование разнообразия и распространения *Tat* LTR-ретротранспозонов в геномах растений.
3. Сравнительный анализ структурных характеристик найденных элементов.
4. Реконструкция филогенетических взаимоотношений между основными структурными вариантам *Tat* LTR-ретротранспозонов.
5. Разработка более детального сценария происхождения дополнительного домена RNH и его роли в эволюции *Tat* LTR-ретротранспозонов.

Научная новизна работы.

В рамках работы разработан DARTS - алгоритм биоинформатического поиска мобильных элементов, содержащих белок-кодирующие домены. Алгоритм реализует поиск LTR-ретротранспозонов групп Ty1/Copia, Ty3/Gypsy, Bel/Pao, DIRS, а также ретровирусов и Penelope-подобных элементов.

В рамках работы впервые произведён поиск LTR-ретротранспозонов зелёных растений (кластер *Tat*), содержащих добавочный домен aRNH, среди геномов стрептофитовых водорослей, печёночных и антоцеротовых мхов, папоротников и древних таксонов семенных, таких как гингковые, саговниковые, гнетовые, кипарисовые, тисовые, амборелловые, нимфовые и магнолииды.

Всего в рамках поиска было исследовано 94 генома зелёных растений, выявлено пять структур элементов *Tat*, различающиеся по положению добавочного домена aRNH. Две из них описаны впервые.

Данные филогенетического и структурного анализов свидетельствуют в пользу конвергентных процессов с ретровирусами позвоночных и LTR-ретротрансопзонами Chronos и Archon оомицетов. Разнообразие выявленных структур и их взаимоотношения свидетельствуют в пользу единичного события захвата домена aRNH, опровергая прежнюю гипотезу о множественном захвате в ходе эволюции *Tat* LTR-ретротранспозонов.

Изучен процесс деградации нативного домена RNH во всех группах в составе *Tat*. Показана связь между наличием добавочного домена aRNH и деградацией нативного RNH, происходящая во всех группах в равной степени. Деградация сразу двух из трёх доменов семейства RNH в одной из впервые выявленных структур косвенно свидетельствует в пользу деградации как следствия конкуренции между доменами с одной функцией.

Теоретическая и практическая значимость исследования

Результаты работы расширяют представления о ковергентной и модульной эволюции, проливая свет на вопрос о сменяемости и конкуренции модулей. Конвергентное сходство независимо образуемых структур свидетельствует в пользу схожего жизненного цикла для эволюционно удалённых групп мобильных элементов.

Разработанный алгоритм DARTS успешно апробирован для поиска мобильных элементов, специфичных по наличию добавочного домена aRNH. Алгоритм также показал высокий потенциал для поиска других групп белок–кодирующих мобильных элементов (Ty1/Copia, Ty3/Gypsy, Bel/Pao, DIRS, Penelope-подобные элементы) и потенциально может быть адаптирован для поиска non-LTR-ретротранспозонов и ДНК-транспозонов.

Основные положения, выносимые на защиту.

1. Элементы LTR-ретротранспозонов, содержащие добавочный домен рибонуклеазы H (aRNH) возникли в геномах наземных растений после зелёных водорослей, что эволюционно соответствует периоду возникновения дифференцированных тканей.

2. Добавочный домен рибонуклеазы H (aRNH) был приобретен ретротранспозонами растений как результат конвергентных по отношению к ретровирусам позвоночных процессов в аналогичном положении, а элементы с альтернативными вариантами положения данного домена являются его производными, которые закрепились и размножились впоследствии.

3. Процесс деградации нативного домена рибонуклеазы H (RNH) в LTR-ретротранспозонах растений является следствием приобретения добавочного домена рибонуклеазы H (aRNH) .

Вклад автора.

Все основные научные результаты были получены автором самостоятельно. Алгоритм DARTS разработан и написан автором на языке программирования Python самостоятельно. Для запуска алгоритма использовался кластер Европейского исследовательского института биологии и старения (ERIBA,

Гронинген), доступ к которому был получен от Березикова Е.В., заведующего лабораторией регуляции стволовых клеток и механизмов регенерации.

Материалы для работы был взяты из открытых бесплатных источников.

Апробация работы.

Результаты по теме диссертации были опубликованы в 7 публикациях: трёх тезисах научных конференций и четырёх статьях в рецензируемых зарубежных и отечественных журналах, три из которых входят в перечень ВАК.

Структура и объем работы.

Диссертация состоит из введения, обзора литературы, материалов и методов, результатов, обсуждения результатов, заключения, выводов и списка литературы. Работа изложена на 104 страницах, содержит 17 рисунков, 2 таблицы и 7 приложений.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

В настоящей главе представлен обзор текущей литературы по теме диссертационной работы, на основании которого сформулированы цель и задачи исследования.

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

Разработка алгоритма поиска и извлечения ретротранспозонов.

Самостоятельно разработанный на языке Python алгоритм поиска, извлечения и аннотации ретротранспозонов основывался на следующих биоинформатических инструментах: поиск доменов реализовывался standalone-версией онлайн-инструмента CD-search (основан на алгоритме RPS-BLAST), кластеризация производилась инструментами CD-hit и MMseqs2, выравнивание - программным обеспечением MAFFT, построение филогенетических деревьев - инструментами FastTree и IQtree. Для статистической поддержки максимального правдоподобия в программе IQ-Tree были использованы два коэффициента: «приблизительный» тест максимального правдоподобия (aLRT) и сверхбыстрый бутстреп (UfBoot). Извлечение последовательностей элементов и доменов, трансляция нуклеотидных последовательностей доменов в аминокислотные, выбор элементов-представителей для сформированных кластеров, фильтрация данных производилась самописными скриптами в составе алгоритма, написанными на языке программирования Python с применением пакетов анализа биологических данных Biopython.

Источники геномных данных.

В ходе работы были использованы базы данных plaBi, NCBI Genomes, GigaDB, Dryad, GyDB, FernBase, Phytozome, TreeGenes, JGI, CNGBdb, NGDC для

получения последних версий геномных сборок представителей нецветковых и древних цветковых растений различных таксонов. Для получения двух геномов антоцеротовых мхов использовался сайт Университета Цюриха (https://www.uzh.ch/en.html), на котором они были опубликованы. В базе данных plaBi представлен список растений, для которых имеется прочтённый геном и ссылка на базу данных с ним. Всего было исследовано 94 геномные сборки зелёных растений.

Методы биоинформационного анализа.

В дополнение к методам, использованным в рамках работы разработанного алгоритма, использовались также следующие методы:

Для пространственного выравнивания по трёхмерной структуре использовался онлайн-инструмент PROMALS3D. Для поиска дополнительных открытых рамок считывания и определения кодируемых ими структур использовались онлайн-ресурсы ORFfinder и CD-search.

ГЛАВА 3. РЕЗУЛЬТАТЫ

Разработка и применение алгоритма DARTS

На языке программирования Python был разработан и написан автоматизированный алгоритм поиска и извлечения последовательностей *Tat* LTR-ретротранспозонов и/или других элементов с доменом aRNH с варьирующими параметрами. Алгоритм позволяет выбрать первичный поисковой домен (для данной работы таковыми прописаны 2 домена - aRNH и RT), программу кластеризации (CD-hit или MMseqs2), специфичность поиска по домену RT (только среди Ty3/Gypsy или по всем доступным RT-профилям из CDD). На данный момент алгоритм, который мы назвали DARTS (Domain-Associated RetroTransposon Search), апробирован и опубликован.

Апробация алгоритма была проведена на 4 геномах растений: *Arabidopsis thaliana*, *Nicotiana tabacum*, *Selaginella moellendorffii* и *Zea mays*. Алгоритмом DARTS и алгоритмом поиска, основанном на инструменте LTRharvest, производились два поисковых запроса - элементов *Tat* (содержащих домен aRNH) и всех элементов Ty3/Gypsy. Алгоритм DARTS показал более высокую чувствительность. Численное сравнение количества обнаруженных элементов двух групп представлено на рисунке 1.

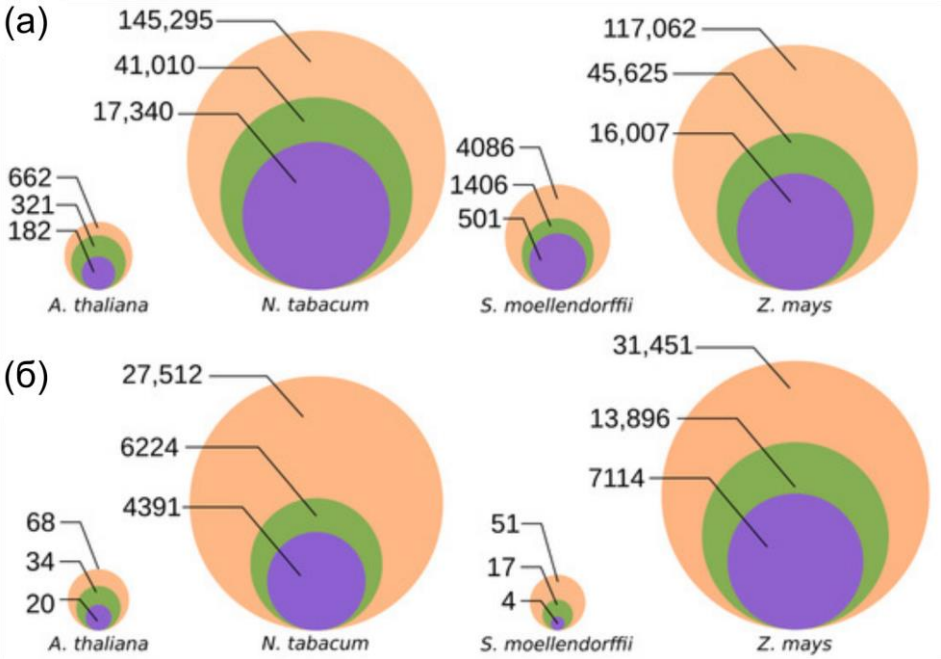

Рисунок 1. Чувствительность идентификации LTR-ретротранспозонов посредством алгоритма DARTS и алгоритма, основанного на инструменте LTRharvest. Оранжевые круги отражают количество элементов, найденных DARTS; зелёные — число элементов, найденных DARTS, содержащих последовательности LTR (потенциальное количество элементов, которые должен был обнаружить LTRharvest); фиолетовые — количество элементов, найденных LTRharvest. Размеры кругов пропорциональны количеству элементов. Количественные значения элементов представлены подписями слева от кругов. Параметры поиска посредством LTRharvest были идентичными для каждого подхода **(а, б)**. **(а)** LTR-ретротранспозоны предсказанные DARTS через поиск по домену RT; элементы, полученные через поиск LTRharvest фильтровали по наличию домена RT. **(б)** LTR-ретротранспозоны, предсказанные DARTS через поиск по домену aRNH; элементы, выявленные как с помощью DARTS, так с помощью LTRharvest фильтровали по одновременному наличию доменов RT и aRNH.

Распространение элементов с aRNH в геномах растений

Всего было найдено около 799 тысяч элементов с доменом aRNH, распознанных DARTS. Распространение выявленных элементов по отделам представлено в таблице 1.

Элементы, содержащие добавочный домен aRNH, не были обнаружены среди всех исследованных геномов зелёных водорослей, а также маршанциевых мхов. Это подтверждает прежние предположения, что процесс приобретения добавочного домена произошёл у растений с уже сформировавшейся дифференцировкой тканей. Было обнаружено как минимум по одной линии элементов с aRNH в геномах настоящих мхов (4 из 8), антоцеротовых мхов (3/3), плаунов (4/4), папоротников (6/9), древних голосеменных - гинкговых (1), саговниковых (1), гнетовых (2/2), тисовых (1/1) и кипарисовых (3/3), современных голосеменных - хвойных (13/13), а также в древних таксонах цветковых - амборелловых (1), кувшинкоцветных (2/2) и магнолиид (6/6). Среди них впервые были обнаружены элементы с aRNH в таксонах: настоящие и антоцеротовые мхи,

папоротники, гинкговые, гнетовые, тисовые и кипарисовые. Прежде считалось, что приобретение добавочного домена произошло на уровне плаунов.

Таблица 1. Результаты поиска LTR-ретротранспозонов методом DARTS, начинающим поиск с обнаружения домена aRNH. Представлено распределение элементов по таксонам, количеству и разнообразию структурной организации.

Таксон	Количество геномов	Количество элементов	Количество и номера структур
Отд. Зелёные водоросли	36	0	0
Отд. Стрептофитовые водоросли	6	0	0
Отд. Маршанциевые мхи	2	0	0
Отд. Настоящие мхи	8	32	1 (#4)
Отд. Антоцеротовые мхи	3	32	2 (#1,#5)
Отд. Плауновидные	4	236	4 (#1,#2,#3#4)
Отд. Папоротникообразные	9	253	1 (#1)
Кл. Гнетовидные	2	27039	2 (#1,#2)
Кл. Гинкговые	1	21357	3 (#1,#2,#3)
Кл. Саговниковые	1	2922	2 (#2,#3)
Кл. Голосеменные, Conifers II	4	343330	2 (#1,#2)
Кл. Голосеменные, Conifers I	9	372339	2 (#2,#3)
Кл. Покрытосеменные	9	31343	1 (#1)

Структурное разнообразие LTR-ретротранспозонов с доменом aRNH в геномах наземных зеленых растений

Все обнаруженные в рамках данной работы элементы были представлены пятью структурами, отличающимися положением домена aRNH относительно других доменов. Структуры и таксоны, к которым они принадлежат, отражены на рисунке 2, а в таблице 1 отражено количество структур, выявленных в каждом таксоне.

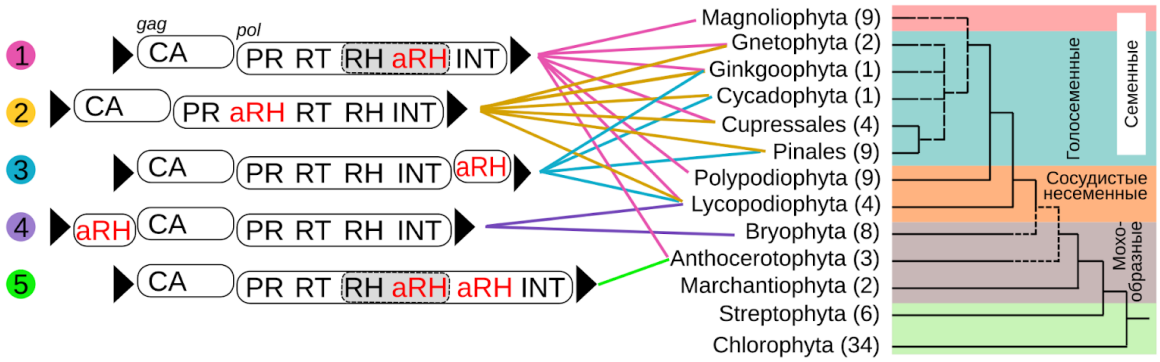

Рисунок 2. Схематичное изображение пяти доменных структур элементов Tat, различающихся по положению домена aRNH, и таксонов растений, в которых они были выявлены. Нумерация структур на изображении (1-5) соответствует дальнейшей нумерации в тексте (#1-#5). Схематическое представление филогении основных таксонов растений, представленное справа, базируется на серии публикаций [143–146]. Цветные линии соединяют структуру с таксонами, в которых она была выявленеа. GAG - соответствует всему белку gag (на самом деле состоит из 3 доменов, в данной работе обозначается как единый домен для простоты

описания), PR - протеаза, RT - обратная транскриптаза, INT - интеграза, RH или RNH - рибонуклеаза H, aRH или aRNH - дополнительный домен рибонуклеазы H.

Обнаруженные элементы подтвердили существование трёх структур, выявленных в прежних работах. Две структуры (#4 и #5) были обнаружены впервые.

Структура #1 представлена во всех таксонах, где в элементах *Tat* имеется добавочный домен aRNH, что позволяет предполагать её как эволюционно наиболее выгодную. Отсутствие литературных данных об остальных структурах в других группах LTR-ретротранспозонов и ретровирусов, захвативших добавочный домен RNH также косвенно свидетельствует в пользу данного предположения. Структура #2 в рамках выборки геномов растений в данной работе показала себя самой многочисленной, копийность которой выше, чем у структур #1 и #3 в геномах голосеменных, являясь таким образом второй по распространенности среди выявленных структур. Это позволяет считать её конкурентоспособной относительно структуры #1. В работе оспариваются предыдущие сообщения о положении добавочного домена aRNH в составе гена *pol* в структуре #3, поскольку выявлено множество случаев расположения домена в отдельной открытой рамке считывания (ОРС) после *pol*. Структура #4 аналогична структуре #3, но дополнительная ОРС с доменом aRNH расположена перед геном *gag*. Структура #5, отличающаяся от остальных не положением, а количеством доменов семейства RNH, может быть рассмотрена как производная от #1, и, наоборот. Поскольку это единственная структура с тремя доменами RNH, её обнаружение приводит к новым вопросам в эволюции *Tat*, например, о её возникновении.

Кластеры элементов *Tat* на основе филогении по домену RT

Инструментом IQTree была проведена филогенетическая реконструкция по трём основным доменам (RT, INT и нативному RNH). Полученные группы на деревьях соответствовали друг другу за исключением незначительных пертурбаций. Всего было обнаружено 12 кластеров (дальнейшая нумерация - *Tat A-L*) в составе *Tat*, содержащих домен aRNH, и 1 базальный кластер (*Tat Z*) без приобретённого добавочного домена. Схема филогенетического древа с подписанными структурами и таксонами, к которым относились данные элементы, представлена на рисунке 3.

Обнаружение кластеров *Tat I, J* и *K,* представленных геномами древних таксонов голосеменных опровергает прежние предположения, что структура #1 сформировалась после дивергенции семенных растений на голосеменных и цветковых.

Присутствие всех обнаруженных структур в ранних таксонах растений (сосудистых несеменных) позволяет выдвигать гипотезу о том, что всё многообразие структур *Tat* возникло из некой первично захватившей домен aRNH

структуры на заре эволюции элементов *Tat* с добавочным доменом. Данные структуры затем элиминировались по-разному от таксона к таксону. Наличие структуры #1 практически во всех исследованных таксонах позволяет предполагать её либо как первично возникшую, либо как эволюционно более выгодную, поскольку именно она сохраняется в большинстве таксонов.

В силу неопределённости происхождения структурного разнообразия, удалённость кластеров *Tat I-L* от *Tat A-D* может рассматриваться как значительно дивергировавшие от единого общего предка-элемента со структурой #1, прошедшие огромное эволюционное расстояние. Аналогичное предположение возникает при рассмотрении группы кластеров *Tat G-H* и кластера *Tat C*. Нам видится более вероятным для каждой группы предположение о дивергенции от общего предкового элемента с соответствующей структурой, возникших в ранних *Tat* после образования всего структурного разнообразия.

Образование в кластере *Tat J* элементов со структурой #2, вероятно, является результатом недавней рекомбинации в силу очень высокой идентичности последовательностей доменов RT для элементов тисовых обеих структур (#1, *Tat I*, и #2, *Tat J*).

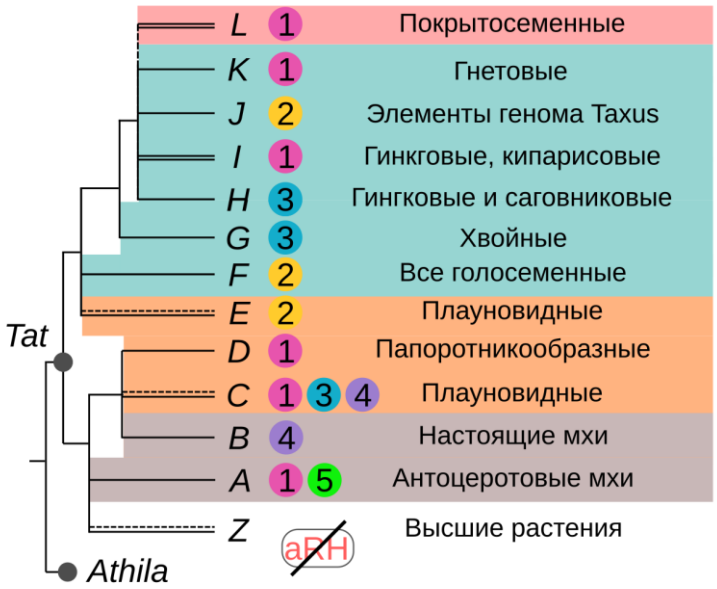

Рисунок 3. Схематичное филогенетическое древо взаимоотношений *Tat* LTR-ретротранспозонов, построенное методом максимального правдоподобия на основе аминокислотных конкатенированных последовательностей доменов RT, RNH, INT, и выполненное в виде кладограммы с условными длинами ветвей. Консенсусные доменные структуры элементов, входящих в каждый кластер обозначены справа от древа. Древо укоренено последовательностями элементов *Chromoviridae*, не отражёнными на рисунке. Следом за следующей внешней группой (близкородственный к *Tat* кластер i) ответвляются кластеры i, обозначенные латинскими буквами *A-L* и *Z*. Кластер *Tat* делится на две ветви (*A-D+Z* и *E-L*). Первое ответвление одной из двух основных ветвей ***Tat***, кластер ***Z*** - кластер, предположительно наиболее древний в *Tat*, имеет структуру, аналогичную внешней группе (не содержит домена aRNH). ***A*** - элементы антоцеротовых мхов со структурами #1 и #5; ***B*** - элементы настоящих мхов со структурой #4; ***C*** - элементы плаунов со структурами #1, #3 и #4; ***D*** -элементы папоротников со структурой #1; ***E*** -

элементы плаунов со структурой #2; ***F*** - кластер общий для элементов из всех групп голосеменных со структурой #2; ***G*** - кластер элементов из хвойных голосеменных со структурой #3; ***H*** - элементы древних голосеменных (гинкговые и саговниковые) со структурой #3; ***I*** - элементы древних голосеменных (гинкговые и кипарисовые) со структурой #1; ***J*** - элементы тисового генома с повторно образовавшейся структурой #2; ***K*** - элементы древних голосеменных (гнетовые) со структурой #1; ***L*** - элементы покрытосеменных со структурой #1. Справа от структур расположен столбец с нумерацией структур (1-5 в цветных кругах соответствуют #1-5) и таксоны, в которых данные элементы были выявлены. Для кластера *Tat Z* таксономический состав обозначен как “Высшие растения”, что означает присутствие элементов данной группы во всех таксонах от трёх групп мохообразных (Настоящие, антоцеротовые и печёночные мхи) до голосеменных и покрытосеменных, исследованных в данной работе.

Филогенетические взаимоотношения доменов aRNH кластеров *Tat*

На полученном из выравнивания по третичной структуре после филогенетической реконструкции древе доменов RNH можно выделить 3 глобальные ветви: (1) ветвь нативных доменов RNH ретротранспозонов, (2) ветвь генов RNH грибов, животных и доменов RNH ретровирусов, (3) ветвь доменов aRNH, куда входят соответствующие гены растений, архей и страменопил. Схематическое изображение филогенетической реконструкции по домену RNH (с двумя остальными ветвями в качестве внешних групп) изображено на рисунке 4. В глобальном плане, оно повторяет филогенетическое древо доменов RNH из прошлого исследования, т.к. содержит те же три основные ветви.

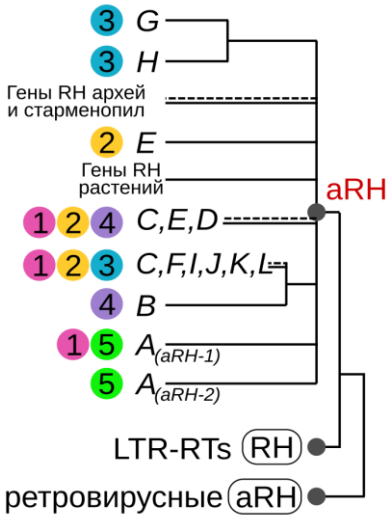

Рисунок 4. Схематическая визуализация филогенетического древа доменов RNH с акцентом на кластера доменов aRNH, выявленных преимущественно в элементах *Tat* (узлы, не имевшие достоверных значений, объединены до ближайших базально расположенных к ним достоверных узлов; длины ветвей условны). Среди трёх основных эволюционных ветвей домена рибонуклеазы H (**ретровирусные aRH** - добавочные RNH ретровирусов, к которым также, согласно прежним работам [9,118], попадают свойственные грибам и животным fmRNH - Fungi/Metazoa RNH), **LTR-RTs RH** - RNH нативных доменов LTR-ретротранспозонов, **aRH** - гены RNH растений, страменопил и архей). Оригинальное древо укоренено кластером доменов ретровирусных aRH. В кластере **aRH** наблюдается 6 кластеров добавочных доменов aRNH из элементов *Tat*: (1) новый

кластер aRNH-2 структуры #5 антоцеротовых мхов, (2) новый кластер доменов aRNH структуры #1 и aRNH-1 структуры #5 из геномов антоцеротовых мхов, (3) кластер aRNH семенных растений, плаунов и настоящих мхов, где мхи формируют небольшую базальную ветвь, а семенные растения с одной линией элементов плауна S.kraussiana сформировали за счёт aRNH элементов с #3 (плаунов), #1 и #2 (семенных растений) структурами основную часть ветви (домены элементов #1 и #2 структур внутри этого кластера не могут быть разделены на внутренние линии), (4) полифилитичный кластер aRNH плаунов различных положений и aRNH папоротников с их единственной структурой, (5) ещё один кластер из одной линии элементов плаунов из генома I.engelmannii со структурой #2, (6) aRNH элементов со структурой #3, состоящий из двух достоверно разделяемых подкластеров - элементов хвойных голосеменных и элементов голосеменных из саговниковых и гинкговых. Обозначения *A-L* соответствуют филогенетическим кластерам по древу RT из рисунка 3. Кластеры выделены цветами по соответствию структурам на рисунке 2: розовый - #1, желтый - #2, синий - #3, фиолетовый - #4 и зелёный - #5.

Все нативные домены RNH, как ожидалось, попали к остальным нативным доменам RNH LTR-ретротранспозонов, сформировав на древе кластер *Tat*, практически аналогичный по структуре древу по домену RT, включая *Tat Z*. Все домены aRNH из обнаруженных алгоритмом элементов *Tat* попали, как и ожидалось, в ветвь aRNH, сформировав внутри неё 6 кластеров (Рисунок 4). Взаимоотношение этих кластеров между собой остаётся спорным, поскольку многие узлы не прошли выбранного критерия достоверности (поддержка aLRT и ufBoot - 80 и 95, соответственно). Низкие коэффициенты статистической поддержки можно объяснить небольшой длиной относительно консервативного участка аминокислотной последовательности доменов RNH, а также быстрым процессом дивергенции этих доменов от общего предка, сопряженным с высокой скоростью мутированиясамих ретротранспозонов.

Филогения по домену aRNH (Рисунок 4) во многом соответствует филогении по домену RT (Рисунок 3) - большинство кластеров остаются таксон-специфичными в соответствии с древом RT, однако, порядок расположения кластеров нарушен.

Сравнение степени деградации нативного домена RNH после приобретения aRNH

В ходе рассмотрения выравнивания аминокислотных последовательностей доменов RNH (aRNH и нативных RNH) с учетом предсказания третичной структуры инструментом PROMALS3D было показано, что нативный домен теряет от двух до четырех аминокислот каталитически активного центра тетрады DEDD, если есть домен aRNH в любом положении. Схема, отражающая целостность активных центров двух доменов для каждой группы кластер-структура-таксон представлена на рисунке 5.

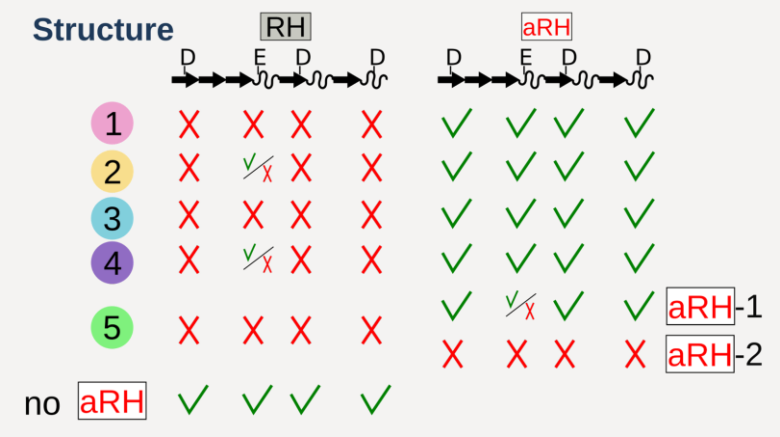

Рисунок 5. Зависимость деградации нативного домена RNH от наличия домена aRNH в различных положениях. Столбец “Кластер” представлен вместе с филогенетическим древом по домену RT из рисунка 3. Столбец “Таксон” отражает таксономический состав групп (за исключением приписки в антоцеротовых мхах, относящейся к “табличной” правой части изображения). Столбец “Структура” отражает, какая из структур (#1-#5) или отсутствие таковой свойственно данной связке кластер-таксон. Справа от столбца “Структура” в табличном виде изображены 8 столбцов по 4 на каждый домен (нативный RNH и aRNH), отражающих наличие в последовательности элемента на соответствующих позициях аминокислот каталитически активного центра тетрады DEDD. На всём рисунке домены обозначены сокращённо как “RH” и “aRH”, соответственно.

Основным наблюдением в данном вопросе является на данный момент связь между деградацией нативного домена RNH с приобретением дополнительного домена aRNH. Во всех группах с aRNH нативный домен RNH потерял хотя бы 2 (в большинстве случаев 3 или 4) аминокислоты тетрады DEDD активного центра. В группах без aRNH нативный домен целый, и, видимо, сохраняет активность.

Также небезынтересно наблюдение, что элементы антоцеротовых мхов со структурой #5, где присутствуют 3 домена RNH (нативный RNH и два aRNH) также содержат лишь один целый домен aRNH-1, в то время как нативный RNH и избыточный aRNH-2 потеряли все аминокислоты активного центра.

Вышеописанные наблюдения подтверждают предшествовавшие данные о деградации нативного домена в *Tat*, показывая тенденцию, аналогичную той, что произошла в ретровирусах. “Старый” домен RNH постепенно теряет аминокислоты каталитически активного центра и, в случае “ретровирусного” (структура #1) положения aRNH имеет перспективы превращения в домен-связку, как это произошло у ретровирусов позвоночных и LTR-ретротранспозонов оомицетов.

Не ясно также, влияет ли на деградацию нативного домена в направлении домена-связки конкретное положение домена aRNH или сам факт его присутствия. В случае ретровирусов позвоночных и LTR-ретротранспозонов оомицетов *Chronos* и *Archon* были обнаружены лишь “ретровирусные” структуры. Это может быть объяснено как тем, что (1) захват происходил в данном положении и далее перемещения никуда не происходило, так и тем, что (2) свидетельства о существовании в данных группах положений aRNH, отличных от

“ретровирусного”, просто не дошли до наших дней. В пользу гипотезы о том, что нативный домен деградирует в домен-связку из-за присутствия другого модуля с тем же свойством (домен aRNH) косвенно может говорить деградация домена aRNH-2 в трёхоменной структуры #5 антоцеротовых мхов. По всей видимости, деградация происходит из-за конкуренции доменов, и один оказывается невостребованным.

ГЛАВА 4. ОБСУЖДЕНИЕ

В данной главе производилось сравнение полученных результатов с предшествовавшими литературными данными, а также выдвигались и разбирались гипотезы, объясняющие полученные результаты.

Предполагаемый эволюционный сценарий

В ходе эволюции элементов *Tat* имеет смысл выделить 4 ключевых события: Захват добавочного домена aRNH, деградация нативного домена RNH, рекомбинационные процессы, сформировавшие структурное многообразие и постепенная элиминация структур.

Мы предполагаем, что первостепенно должны были произойти события захвата и деградации, поскольку во всех структурах, независимо от таксона, деградация нативного домена RNH выявляется в равной степени. Оба этих события должны были иметь место у организма, являющегося общим предком для всех современных таксонов, в которых элементы *Tat* с доменом aRNH были обнаружены. Так, мы предполагаем, что это происходило в геноме некоего раннего предка всех групп мхов и плаунов. Нет возможности однозначно утверждать, какой из этих двух процессов следовал за другим. Однако, в силу свидетельств о конкуренции между доменами и отсутствия обнаруженных элементов без aRNH с деградировавшим RNH, мы склоняемся в пользу гипотезы, где деградация нативного RNH следует за захватом aRNH.

Захват домена aRNH следует считать событием единократным. Филогенетическая реконструкция не даёт возможности однозначно отказаться от гипотезы о множественном захвате, однако, 6 или 7 случаев захвата одного домена из крайне близких источников на уровне одного организма видится нам сценарием менее вероятным. В результате захвата сформировалась некая первичная структура, положение домена aRNH в которой остаётся неизвестным. Мы допускаем, что этой структурой может быть структура #1 в силу её наибольшей распространённости. Однако, последовавшая серия рекомбинационных процессов, породившая все остальные структуры, могла, и, вероятно, породила всё многообразие структур в пределах одного генома. А потому структурой-прародительницей может являться любая другая, в том числе не сохранившаяся до наших дней структура. В целом, процесс рекомбинации привёл к формированию как минимум 5 структур, 4 из

которых сохранились на уровне одного из наиболее древних из из исследованных таксонов - плаунов.

Мы выдвигаем гипотезу о постепенной таксон-специфической элиминации структур в ходе дивергенции и специализации таксонов зелёных растений. Альтернативной гипотезой, объяснившей бы подобное многообразие, могло бы послужить предположение, что структуры постоянно рекомбинировали между собой, а геномы активно обменивались ими в результате горизонтального переноса. В силу практически однозначно показанной филогенетической реконструкцией по домену RT, вертикальной эволюции элементов *Tat*, подобный сценарий является крайне маловероятным.

ГЛАВА 5. ЗАКЛЮЧЕНИЕ

Проанализировано 94 генома зелёных растений. Элементы, содержащие домен aRNH, были обнаружены в геномах настоящих и антоцеротовых мхов, отбрасывая гипотезу о том, что захват aRNH элементами *Tat* произошёл в геномах плаунов (список исследованных таксонов с выявленными элементами и их численностью представлен в таблице 1). Для выявленных элементов *Tat* с aRNH и сформированных ими кластеров внутри *Tat* было введено новое условное обозначение (*A*-*L* вместо I-VI). Всего было выявлено около 799 тысяч элементов с aRNH, филогенетически соответствующих кластеру *Tat*. Эти элементы были представлены пятью различающимися по положению домена aRNH структурами, две из которых (#4 и #5) были выявлены впервые (Рисунок 2).

LTR-ретротранспозоны Tat с доменом aRNH из геномов папоротников, плаунов, настоящих и антоцеротовых мхов сформировали на древе RT (рисунок 3) четыре новых кластера *(Tat A*, *B*, *C*, *D*) за пределами “старых” кластеров *Tat* (по новой нумерации - *E*, *F*, *G* и *L*). Ещё четыре кластера (*Tat H*, *I*, *J* и *K*) были выявлены в древних голосеменных (гингкового *Ginkgo biloba*, гнетовых *Gnetum montanum* и *Welwitschia mirabilis*, *Cycas panzhihuaensis*, тисового *Taxus wallichiana* и трёх видов семейства кипарисовых), расположенный между известными прежде кластерами *Tat G* и *L* (хвойных голосеменных и покрытосеменных, соответственно).

Элементы с “ретровирусной” структурой #1 были выявлены у папоротников, плаунов и антоцеротовых мхов и древних голосеменных и покрытосеменных, что делает эту структуру самой распространённой и, вероятно, эволюционно наиболее выгодной. Нами рассматриваются две гипотезы. Согласно одной, захват aRNH элементами *Tat* произошёл в положение структуры #1, от неё же произошли все остальные структуры. Это объясняет представленность структуры #1 практически во всех исследованных таксонах, содержащих элементы Tat с доменом aRNH. Согласно альтернативной гипотезе, захват мог произойти в любом другом положении, однако, только после образования структуры #1 в ходе рекомбинации

элементы Tat с доменом aRNH получили эволюционное преимущество, закрепившись во всех таксонах преимущественно в виде “ретровирусной” структуры (#1). В любом случае, образование данной структуры в ходе эволюции LTR-ретротранспозонов в геномах зелёных растений мы расцениваем как пример конвергенции по отношению к ретровирусам позвоночных и LTR-ретротранспозонам оомицетов.

Аналогичные конвергентные процессы на уровне разнообразия внутри кластеров *Tat* показаны на двух других структурах - #2 (кластеры *Tat E-F* и *Tat J*) и #3 (*Tat C* и *Tat G-H*).

Новая структура #4, где домен aRNH расположен в 5` области до ОРС гена *gag*, выявлена в геномах настоящих мхов и плаунов. Новая структура #5 с двумя доменами aRNH, вероятно, являющаяся производной структуры #1 (“ретровирусной”), обнаружена в геномах антоцеротовых мхов (Рисунки 2, 3).

Филогенетический анализ домена aRNH (Рисунок 4) предполагает единственный случай приобретения домена aRNH у *Tat* LTR-ретротранспозонов с последующей дивергенцией.

Было проведено исследование зависимости деградации нативного домена рибонуклеазы H от приобретения дополнительного домена aRNH (Рисунок 5). Связь этих двух событий была подтверждена, однако определить последовательность, в которой эти события происходили, остаётся невозможным. В силу особенностей выявленной структуры #5 (Рисунок 2), мы склоняемся к тому, что процесс деградации старого домена является скорее следствием приобретения нового, конкурирующего с ним.

Также был произведён ограниченный поиск добавочных ОРС, в ходе которого выявлены некоторые добавочные фрагментированные домены, функциональная роль которых остаётся неясной.

Согласно логике рассуждения в прежних работах, мы должны были бы утверждать о 6 или 7 случаях независимого захвата домена aRNH. Однако в силу пересечений между этими кластерами, вероятно, многие из них являются образовавшимися из одного источника путём дивергенции. В силу малой длины консервативной последовательности домена aRNH, филогенетическое древо не позволяет разрешить взаимоотношения кластеров в пользу множественных событий захвата домена aRNH. Следовательно, гипотеза о единственном случае захвата видится более вероятной. Таким образом, мы предлагаем отказаться от ранее выдвинутой нами гипотезы о неоднократных независимых случаях приобретения домена aRNH элементами кластера *Tat*.

Поскольку не представляется возможным выяснить, как часто происходили процессы рекомбинации между структурами *Tat* по положению домена aRNH, нет

возможности однозначно утверждать, является ли выявленное разнообразие элементов *Tat* результатом постоянных рекомбинаций или же постепенной утратой разными таксонами разных структур, образовавшихся сразу после захвата домена aRNH.

ГЛАВА 6. ВЫВОДЫ

1. Разработан автоматизированный алгоритм биоинформационного поиска, извлечения, аннотации и классификации последовательностей LTR-ретротранспозонов с дополнительным доменом aRNH, а также других мобильных элементов с консервативными белковыми доменами. Чувствительность алгоритма превышает ранее разработанные методы.
2. Определено распространение *Tat* LTR-ретротранспозонов в геномах нецветковых зеленых растений. Элементы *Tat* были выявлены впервые у представителей таксонов настоящих и антоцеротовых мхов, папоротников и древних групп голосеменных - гинкговых, гнетовых, саговниковых, тисовых и кипарисовых.
3. Приобретение домена aRNH *Tat* LTR-ретротранспозонами произошло на уровне предка наземных сосудистых растений после дивергенции от зеленых водорослей, а не на уровне плаунов, как это считалось ранее. Далее, эволюция и распространение *Tat* LTR-ретротранспозонов в геномах растений проходила преимущественно вертикально.
4. Было выявлено 5 структурных вариантов элементов *Tat*, различающихся по положению домена aRNH. Две структуры обнаружены впервые. У *Tat* LTR-ретротранспозонов из большинства изученных таксонов преобладает структура “ретровирусного” типа, что говорит в пользу её приспособительного значения для эволюции LTR-ретротранспозонов, а также первичного происхождения данной структуры у общего предка всех *Tat*. Само по себе возникновение “ретровирусной” структуры у LTR-ретротранспозонов зеленых растений представляет яркий пример конвергентной эволюции по отношению к ретровирусам позвоночных и LTR-ретротранспозонов из геномов паразитических протист - оомицетов.
5. Филогенетическая реконструкция отношений последовательностей генов и доменов RNH подтверждает, что aRNH происходит не от нативного домена RNH, а путём однократного приобретения aRNH транспозоном извне с последующей быстрой дивергенцией домена. Таким образом, отвергнута предыдущая гипотеза о множественных независимых приобретениях aRNH различными кластерами *Tat*.
6. Деградация нативного домена RNH во всех кластерах *Tat*, содержащих aRNH в различных положениях, происходит в равной мере для всех изученных групп.

Это подтверждает взаимосвязь двух событий: приобретение aRNH и деградация RNH, но не позволяет явно определить их последовательность. Пример исследования дуплицированных доменов aRNH в структуре #5 элементов *Tat* антоцеротовых мхов позволяет с большей вероятностью допустить, что деградация нативной RNH является следствием конкуренции двух или более доменов с одинаковой функцией.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Ustyantsev K., **Biryukov M.**, Sukhikh I., Shatskaya N.V., Fet V., Blinov A., Konopatskaia I.. Diversity of *mariner*-like elements in Orthoptera. // Вавиловский журнал генетики и селекции. 2019;23(8):1059-66.
2. **Biryukov M.**, Ustyantsev K.. DARTS: An Algorithm for Domain-Associated Retrotransposon Search in Genome Assemblies: 1 // Genes. Multidisciplinary Digital Publishing Institute, 2022;13(1):9.
3. Martinez P., Ustyantsev K., **Biryukov M.**, Mouton S., Glasenburg L., Sprecher SG, Bailly X., Berezikov E. Genome assembly of the acoel flatworm Symsagittifera roscoffensis, a model for research on photosymbiosis. // G3 Genes| Genomes| Genetics. 2022; , P. 00(0), jkac336.
4. **Biryukov M.**, Berezikov E., Ustyantsev K. Classification of LTR retrotransposons in the flatworm Macrostomum lignano. Letters to Vavilov Journal of Genetics and Breeding. 2020. 6(2). DOI: 10.18699/Letters2020-6-12.

МАТЕРИАЛЫ И ТЕЗИСЫ КОНФЕРЕНЦИЙ

1. **Biryukov M.**, Ustyantsev K.. Diversity and evolution of *Tat* LTR retrotransposon structures in non-flowering plants. // Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2020). Novosibirsk. P. 201.
2. **Biryukov M.**, Ustyantsev K.. LTR retrotransposons in green plants show multiple examples of convergent evolution. // Systems Biology and Bioinformatics (SBB-2020).
3. **Biryukov M.**, Ustyantsev K.. Diversity and evolution of structural variants of *tat* ltr-retrotransposons in green plants. // EMBO Workshop The Mobile