

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
НАУЧНОЕ УЧРЕЖДЕНИЕ «ФЕДЕРАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИНСТИТУТ ЦИТОЛОГИИ И
ГЕНЕТИКИ СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ
АКАДЕМИИ НАУК»

На правах рукописи

БИРЮКОВ МИХАИЛ ЮРЬЕВИЧ

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

1.5.7. – генетика

Диссертация на соискание ученой степени
кандидата биологических наук

Научный руководитель:
кандидат биологических наук,
н.с. Устьянцев Кирилл Валерьевич

Новосибирск 2023

Оглавление

Список сокращений	4
Введение	6
Актуальность	6
Цель и задачи исследования	7
Научная новизна работы	7
Теоретическая и практическая значимость исследования	8
Основные положения, выносимые на защиту	9
Вклад автора	9
Апробация работы	10
Структура и объем работы	11
Глава 1. Обзор литературы	12
Мобильные генетические элементы: класс ретротранспозоны	12
Вклад мобильных генетических элементов в геномы эукариот и его эволюционные последствия	12
Особенности классификации ретротранспозонов	15
LTR-ретротранспозоны	16
Строение LTR-ретротранспозонов	17
Жизненный цикл LTR-ретротранспозонов	18
Функция домена RNH в LTR-ретротранспозонах	23
Tat LTR-ретротранспозоны	25
Конвергентная эволюция групп ретротранспозонов, имеющих домен aRNH	26
Заключение	29
Глава 2. Методы исследования	31
Источники геномных данных	31
Алгоритм поиска и извлечения Tat LTR-ретротранспозонов	31
Определение зависимости деградации нативного домена RNH от наличия добавочного домена aRNH в различных положениях	39
Поиск дополнительных открытых рамок считывания	40
Глава 3. Результаты	41
Разработка и применение алгоритма DARTS	41
Распространение элементов с aRNH в геномах растений	43
Структурное разнообразие LTR-ретротранспозонов с доменом aRNH в геномах наземных зеленых растений	45
Кластеры элементов Tat на основе филогении по домену RT	47
Филогенетические взаимоотношения доменов aRNH кластеров Tat	52
Сравнение степени деградации нативного домена RNH после приобретения aRNH	55
Анализ спорных элементов-синглетов	56
Поиск добавочных OPC с кодирующими последовательностями	58

Глава 4. Обсуждение результатов	60
Несоответствия деревьев RT и RNН	60
Деградация нативного домена RNН	63
Структуры: найденные, вероятные, производные и преобладающие	65
Специфичность кластеров Tat таксонам растений	69
Предполагаемый эволюционный сценарий	71
Заключение по проделанной работе	74
Выводы	77
Библиографические ссылки	79
Приложения	89

Список сокращений

МГЭ - мобильные генетические элементы

РНК - рибонуклеиновая кислота

ДНК - дезоксирибонуклеиновая кислота

кДНК - комплементарная ДНК

мРНК - матричная РНК

тРНК - транспортная РНК

LTR - long terminal repeat - длинные концевые повторы

RNH - RNase H type I - рибонуклеаза H 1 типа

aRNH - “Archeal”/Archea-like RNH - “архейный”/дополнительный домен RNH

LTR-RNH, LTR RNH - нативная RNH LTR-ретротранспозонов

fmRNH - fungi-metazoa RNH - домен RNH грибов и животных

TSD - Target Site Duplication - дупликация целевого сайта

DARTS - Domain-Associated RetroTransposon Search - домен-ассоциированный поиск ретротранспозонов

млрд - миллиард

п.о. - пары оснований

т.п.о. - тысячи пар оснований

RT - reverse transcriptase - обратная транскриптаза

PR - protease - протеаза

INT - integrase - интеграз

GAG - белок гена оболочки *gag* [в работе для удобства употребляется как домен, однако, на деле, состоит из трёх доменов - CA, NC и MA]

CA -капсидный домен гена *gag*

NC - нуклеокапсидный домен гена *gag*

MA - матричный домен гена *gag*

ENV - домен гена оболочки *env* (envelope - оболочка)

ORF - открытые рамки считывания (ORF - open reading frame)

дОРФ - добавочные ОРФ

ВПЧ - вирусоподобная частица

PBS - prime binding site - сайт связывания праймера

PPT - polypurine tract - полипуриновый тракт

SSA [pathway] - single-strand annealing - одноцепочечный отжиг

IR - inverted repeats - инвертированные повторы

CC/CCDC - домен/мотив coil-coiled

PMD - plant mobile domain - мобильный домен растений

Tr28 - домен транспозазы 28

SMC - мотив сегрегации хромосом

aLRT - approximate likelihood-ratio test - “приблизительный” тест максимального правдоподобия

UfBoot - ultra-fast bootstrap - ультра-быстрый бутстреп

CDD - conserved domain database - база данных консервативных доменов

ПО - программное обеспечение

DB - database - база данных

PDB - protein database - белковая база данных

UTR - untranslated region - нетранслируемая область

Введение

Актуальность

Мобильные генетические элементы (МГЭ) - последовательности генома, способные менять своё положение в нём [1]. Благодаря этому свойству МГЭ являются важными регуляторами структуры и функционирования генома [2]. 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 [3,4]. 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 [5,6].

Среди ретротранспозонов выделяют отдельный большой подкласс элементов с длинными концевыми повторами (LTR - long terminal repeat) - LTR-ретротранспозоны [7]. LTR-ретротранспозоны структурно и эволюционно близки ретровирусам позвоночных животных, а также обладают схожим жизненным циклом [8]. 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*) [9]. Было показано, что у элементов в данном кластере существуют различные структурные варианты [10], 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) [9]. 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.

Цель и задачи исследования

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

Для выполнения данной цели были поставлены следующие **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 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-ретротранспозонами *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 предположения, что деградация является следствием 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 были взяты из открытых бесплатных источников.

Апробация работы

Результаты данной работы были представлены в виде семи публикаций, среди которых три тезиса научных 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)–2020.–С.201-201.
2. **Biryukov M.**, Ustyantsev K. LTR retrotransposons in green plants show multiple examples of convergent evolution // 12th International young scientists school «System Biology and Bioinformatics» (SBB-2020) – 2020. – С. 24-24.
3. **Biryukov M.** , Ustyantsev K. Diversity and evolution of structural variants of tat ltr-retrotransposons in green plants. // EMBO Workshop The Mobile Genome: Genetic and Physiological Impacts of Transposable Elements. 2021.

Также по теме диссертации были опубликованы четыре статьи в рецензируемых зарубежных и отечественных журналах, три из которых (1-3) входят в список ВАК:

1. K. Ustyantsev, **M. Biryukov**, I. Sukhikh, N.V. Shatskaya, V. Fet, A. Blinov, I. Konopatskaia. Diversity of mariner-like elements in Orthoptera. // Вавиловский журнал генетики и селекции. 2019;23(8):1059-66.
2. **M. Biryukov**, K. Ustyantsev. 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.

Структура и объем работы

Диссертация состоит из введения, обзора литературы, материалов и методов, результатов, обсуждения результатов, заключения, выводов и списка литературы. Работа изложена на 103 страницах, содержит 17 рисунков, 2 таблицы и 7 приложений.

Глава 1. Обзор литературы

Мобильные генетические элементы: класс ретротранспозоны

Ретротранспозоны - класс мобильных генетических элементов (МГЭ), обладающих способностью к самостоятельному перемещению внутри генома (транспозиции) посредством механизма “копирование и вставка” (“copy-and-paste”) [3,4]. Механизм процесса обычно сравнивается с аналогичным механизмом другого большого класса МГЭ, ДНК-транспозонов, который носит название “вырезание и вставка” (“cut-and-paste”) [11]. Механизм “копирование и вставка” заключается в наличии промежуточной стадии в жизненном цикле ретротранспозонов - копии ретротранспозона в виде мРНК. По данной мРНК посредством трансляции синтезируются закодированные в ретротранспозоне белки, которые затем переводят саму мРНК сначала в комплементарную ей ДНК (кДНК), а затем уже в полноценную двуцепочечную ДНК-копию исходного ретроэлемента, которая встраивается в новый участок генома клетки-хозяина [4–6].

Вклад мобильных генетических элементов в геномы эукариот и его эволюционные последствия

В геномах различных эукариотических организмов ретротранспозоны представлены разным числом копий. Ретротранспозоны некоторых групп могут присутствовать менее, чем в десятке копий, или отсутствовать вообще. В тоже время другие группы представлены множеством копий (вплоть до сотен тысяч на геном), что объясняется активной жизнедеятельностью элементов в прошлом и/или настоящем [6,12,13]. Благодаря своему механизму перемещения ретротранспозоны способны образовывать большое число копий, которые, в свою очередь, могут составлять огромную долю от общего размера генома, в котором они находятся. Наиболее явно это выражается у растений, геном которых во многом увеличен именно за счёт ретротранспозонов [12,14,15]. Так, например, в геноме кукурузы, размер которого составляет примерно 2,3 млрд

п.о., около 70-80% от этого размера приходится на ретротранспозоны [16–18]. У некоторых лилейных (сем. *Liliaceae*) это значение доходит до 90% [19].

Эволюционная роль ретротранспозонов, равно как и других МГЭ, неоднозначна. Так, многообразие копий ретротранспозонов служит сайтами для неравного кроссинговера и других хромосомных перестроек, включая гомологические рекомбинации, сайты которых создаются в том числе в результате транспозиций МГЭ [1,2,20–24]. С другой стороны, были показаны и проявления, основанные на инсерционных событиях [25,26].

Так, первым, наиболее простым следствием встройки элемента является его TSD (Target Site Duplication, дупликация целевого сайта) - сайт, по которому пептиды с интегразной/транспозазной активностью ретровирусов и мобильных элементов встраивают последовательности транспозонной или вирусной ДНК в геном хозяев, специфично распознаётся и удваивается. В случае элементов, которые потом могут быть вырезаны из места встройки (ДНК-транспозоны, другой класс мобильных элементов) дупликация целевого сайта приводит к появлению нескольких лишних нуклеотидов [25], способных как внести некое точечное разнообразие, так и создать сдвиг рамки считывания (frameshift) [27–29]. Однако, ретротранспозоны обычно имеют более длинные последовательности, встройка которых вносит гораздо больший пласт вероятных эффектов. Само встраивание достаточно длинной последовательности, а также возникающие следом за ней перестройки, способны изменить структуру гена: внутри гена может быть образован новый интрон [30,31], экзон [32–34], привносятся новые сайты альтернативного сплайсинга [30,31,35–37], происходит сдвиг рамки считывания [28,29,38], приводящий вплоть до нарушения работы гена (нулевой фенотип) [34,39,40]. Также известны случаи слияния хозяйского гена с доменами мобильных элементов, в результате чего образуются новые химерные гены [41,42]. Для новых копий мобильных элементов показано также и влияние на организм хозяина на эпигенетическом уровне [43,44]: встройки в регуляторной области приносят новые или разрушают старые промоторы и сайты-энхансеры [23,45–49], сайты сайленсинга [21,44], полиаденилирования [42] и метилирования [50–52], меняют структуру промоторной области своей повторенной последовательностью или перенося фрагменты или целые промоторы из других областей генома [43,53,54]. Так, присутствие мобильного

элемента часто снижает экспрессию близлежащих генов [55]. Встройка в локусе цветения C (flowering locus C, FLC) у *Arabidopsis thaliana*, напротив, приводит к раннему созреванию и изменению эффективности яровизации [56]. Промоторы также могут быть сформированы самой последовательностью ретротранспозона на примере промоторной области растительного гена LAT59 [57].

Сама активность МГЭ тоже способна внести свои коррективы в клеточные процессы. Активируясь под воздействием различных стрессорных факторов (таких как нагревание [28,58,59] или гамма-радиация [59–62]) продукты жизнедеятельности МГЭ принимают участие в клеточном ответе. Отдельно стоит выделить ретродупликации - встройки в геном процессированных мРНК за счёт ферментов МГЭ (обратная транскриптаза интегразы и пр.) [25,49]. Отдельные группы non-LTR ретротранспозонов были сформированы именно этим путём [3,63].

Особенно важным оказывается тот генетический материал, который может быть перенесён ретротранспозоном вместе с последовательностью самого элемента. Горизонтальный перенос генов позволяет организмам приобретать свойства, не выработанные или утраченные “вертикальным” путём [64–66]. Так, известны случаи захвата различных бактериальных белков геномами папоротникообразных [67,68] и покрытосеменных [67], горизонтального переноса из артропод в покрытосеменные [68] за растениями-паразитами задокументирован горизонтальный перенос генов от их жертв [69,70].

Известны также многочисленные примеры одомашнивания ретротранспозонов или их отдельных генов и доменов, где вместо горизонтально перенесённых генов новые свойства возникают из самих мобильных элементов и провирусов. Так, ретроэлементы *HetA* и *TART* встраиваются в район теломерных концов у дрозофил, не имеющих рабочей теломеразы, и выполняют таким образом функцию этого фермента. Филогенетический анализ также показывает, что сама теломераза произошла от обратной транскриптазы древнего ретротранспозона [71,72]. От ретровирусов ведут своё происхождение прото-онкогенные белки (вирусные киназы) [73–75]. У растений аналогичная тенденция с преобразованием генов LTR-ретротранспозонов [76]. У ретровирусов позвоночных животных, дочерней группы ретротранспозонов, ген оболочки *env* (envelope - оболочка) также был “одомашнен” геномами

млекопитающих и выполняет роль фактора формирования плаценты [77,78]. Разнообразие групп внутри класса ретротранспозонов позволяет наблюдать различные эволюционные приспособления ретротранспозонов, такие как приобретение предком ретровирусов позвоночных гена оболочки, позволившего им выживать во внешней среде и эффективно распространяться.

Особенности классификации ретротранспозонов

Многообразие ретротранспозонов и их групп, как и других групп мобильных элементов, пополняется постоянно. И, хотя некоторое время назад была обнаружена базальная группа ретротранспозонов (PLE, Penelope-подобные элементы), не подходящая под данную классификацию [79–81], ретротранспозоны классифицируют в первую очередь по наличию последовательностей LTR - длинных концевых повторов (long terminal repeat). Так, помимо PLE, выделяют два подкласса: LTR- и non-LTR-ретротранспозоны. Последовательности LTR, как можно судить по их названию, фланкируют ретротранспозон, определяя его границы. Сами по себе LTR не содержат транслируемых рамок считывания, но являются повторяющимися последовательностями с высоким уровнем идентичности [7].

Хотя ключевое отличие подкласса LTR-ретротранспозонов от подкласса non-LTR-ретротранспозонов - это наличие у элементов первого класса LTR, - их также можно различить по последовательности гена обратной транскриптазы (или “ревертазы”, reverse transcriptase, RT) [4]. Этот фермент обеспечивает перевод мРНК в ДНК. Он впервые был обнаружен у ретровирусов позвоночных животных и сначала воспринимался как особенность, свойственная сугубо данной группе. В дальнейшем было показано, что большинство эукариотических организмов содержат ген обратной транскриптазы в количествах, численно превосходящих копии других генов. Как оказалось в ходе в дальнейшем проведенных исследований геномных последовательностей, все они входили в состав ретротранспозонов, создавших внутри эукариотических геномов множество собственных дубликатов [7,82,83]. Сейчас известное многообразие ретротранспозонов значительно превосходит биоразнообразие ретровирусов. Более того, ретровирусы принято считать дочерней группой ретротранспозонов [8,9]. Это можно проследить по эволюции гена RT, поскольку этому гену

характерна эволюционная монофилия в пределах всего класса ретротранспозонов [4]. С помощью гена RT были установлены взаимоотношения различных элементов внутри класса (Рисунок 1).

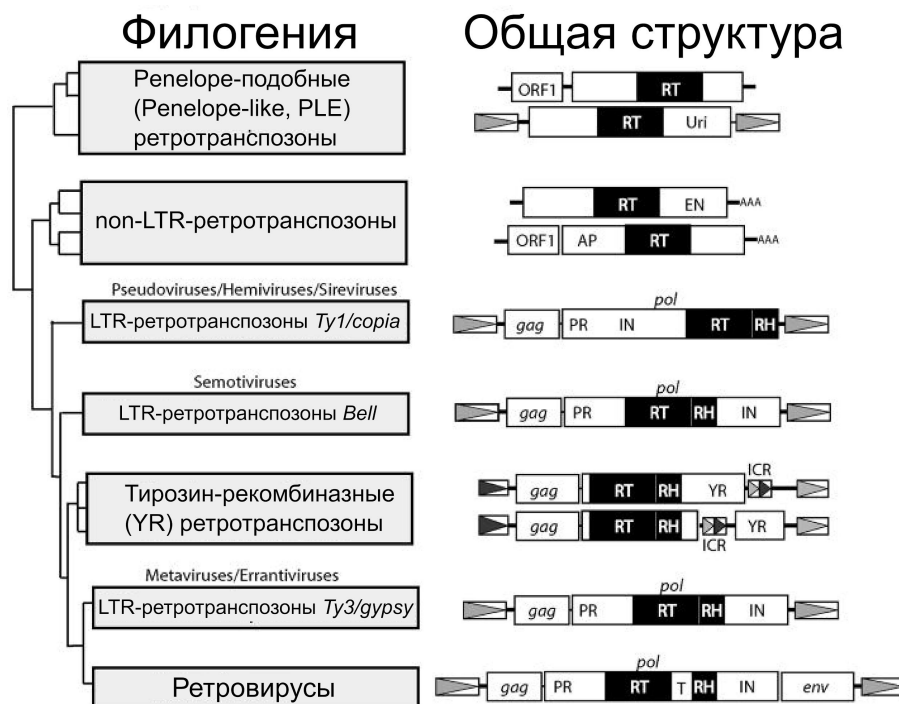


Рисунок 1. Филогения ретротранспозонов, построенная по гену обратной транскриптазы (RT), и общая структура для каждой группы. Над некоторыми группами подписана их классификация по вирусологической номенклатуре. В состав большинства групп входят открытые рамки считывания (ОРС, ORF - open reading frame) *gag*, *pol*, у ретровирусов ещё *env*. Основные домены: PR - аспартильная протеаза, RT - обратная транскриптаза, RNH - рибонуклеаза H, INT - интеграз, YR - тирозин-рекомбиназа, T - “tether” - домен-“связка”, EN - эндонуклеаза, AP - Апуриновая/апиримидиновая эндонуклеаза. Доменные структуры и роль доменов будет разобрано в последующих разделах. Адаптировано из [4].

LTR-ретротранспозоны

Как было сказано выше, LTR-ретротранспозоны характеризуются фланкирующими с каждого края элемента последовательностями LTR. LTR не кодирует белков, однако, в своём составе они содержат промоторный и терминаторный сайты, необходимые для обратной транскрипции [7]. Именно этот подкласс ретротранспозонов является наиболее представленным в геномах растений. Однако вклад в структуру геномов растений реализуется сравнительно малым разнообразием LTR-ретротранспозонов с высоким числом копий. По сравнению с растениями, в геномах животных каждый LTR-ретротранспозон

представлен не более чем десятком копий, но разнообразие элементов на геном шире [6,12,14,15].

Наиболее крупными и детально изученными группами LTR-ретротранспозонов являются группы *Ty3/gypsy* и *Ty1/copia* (Рисунок 1). Их представители также раньше рассматривались как ретровирусы, и потому по вирусологической номенклатуре имеют названия метавирусы (Metaviridae) и псевдовирусы (Pseudoviridae), соответственно. Основной доменный состав этих групп аналогичен (рисунок 1), однако классифицируют эти две, а также и другие, менее изученные, группы LTR-ретротранспозонов (*Bel/Pao*, *DIRS* и др.) по составу кодируемых функциональных белковых доменов, их порядку [7,84].

Строение LTR-ретротранспозонов

Основными структурами, составляющими любой LTR-ретротранспозон, являются фланкирующие последовательности LTR и две (или иногда в результате слияния - одна) открытые рамки считывания (ОРС) внутри [7,15,19]. Этими ОРС кодируются два гена: *gag* и *pol* (или *gag-pol* в случае слияния рамок считывания в одну), - экспрессирующихся в разном соотношении [4,84]. Продукты гена *gag* требуются для сборки упаковочной вирусоподобной частицы (ВПЧ), причём их требуется значительно больше, чем продуктов *pol* [19]. На одной молекуле мРНК подобное стехиометрическое соотношение достигается за счёт проскальзывания рибосомы по матрице. Это может быть обусловлено как за счёт короткого повторённого мотива, так и случайного проскальзывания рамки считывания [85,86].

Полипротеин, получаемый из гена *gag*, структурно схож с белком нуклеокапсида оболочки ретровирусов. У ретровирусов он содержит три домена: капсид, нуклеокапсид и матричный домен. Капсид, или полимеризующий домен, участвует в формировании ВПЧ. ВПЧ, вероятно, востребована для создания среды, где только мРНК ретротранспозона подвергнется обратной транскрипции. Нуклеокапсид - домен, ответственный за взаимодействие с мРНК ретровируса или ретротранспозона. Матричный домен присутствует в явном виде у ретровирусов, поскольку ответственен за связь с белком оболочки. У остальных ретротранспозонов матричный домен выражен не явно или отсутствует [87,88].

Полипротеин гена *pol* регулирует жизненный цикл элемента и содержит несколько доменов: PR - аспартильная протеаза, RT - обратная транскриптаза, RNH - рибонуклеаза H и INT - интегразы. PR отвечает за процессинг белка после трансляции. Этот фермент отщепляет собственный домен от полипептида, полученного после трансляции с мРНК гена *pol*, после чего, уже находясь в свободной форме, повторяет аналогичный процесс с доменом INT. Домены RT и RNH остаются единой структурой, которая отвечает за взаимодействие с мРНК, синтез новой ДНК-копии и последующее расщепление мРНК. Домен INT осуществляет встраивание полученной ДНК-копии элемента обратно в геном [87,89].

В случае слияния ОРС генов *gag* и *pol* в одну рамку, нарушается требуемое стехиометрическое соотношение полипротеинов. Поскольку продуктов гена *gag* требуется значительно больше, для формирования ВПЧ, то в данном случае избыточный продукт с гена *pol* расщепляется [19].

Жизненный цикл LTR-ретротранспозонов

В общем виде жизненный цикл ретротранспозонов может быть разбит на 4 стадии: (1) транскрипция и трансляция, в ходе которых сперва с ДНК ретротранспозона считывается мРНК, с которой затем синтезируются полипротеины Gag и Pol (каждый подвергается расщеплению на три белка); (2) образование рибонуклеопротеинового комплекса (которым является ВПЧ) - мРНК связывается с белковыми структурами, полученными на первом шаге; (3) в ходе обратной транскрипции формируется ДНК-копия ретротранспозона; (4) интеграция новой копии, транспортированной рибонуклеопротеиновым комплексом, в новый участок генома с помощью интегразы [90,91] Обобщенный жизненный цикл представлен на рисунке 2.

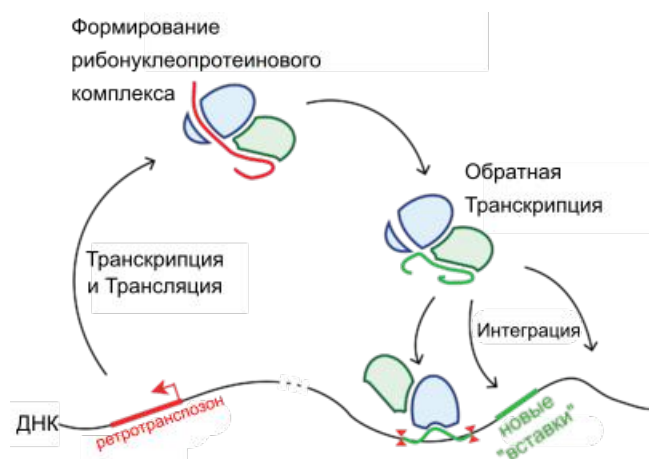


Рисунок 2. Жизненный цикл ретротранспозонов в общем виде, представленный в 4 шага. Адаптировано из [92].

Говоря же об LTR-ретротранспозонах, следует детально разобрать процесс обратной транскрипции, отметив роль, которую играют последовательности LTR. В ВПЧ, состоящей из доменов процессированного продукта гена *gag*, попадают все белковые продукты гена *pol* и мРНК LTR-ретротранспозона. Данная мРНК содержит все ОРС элемента (*gag*, *pol* и дополнительные ОРС, если такие имеются), фланкированные последовательностями LTR. Цепь мРНК модифицирована поли-А последовательностью на 3'-конце. В составе LTR имеется как повторяющаяся последовательность (R), так и уникальные для 5' и 3' LTR последовательности (U5 и U3), соответственно. Между 5' LTR и первой ОРС (гена *gag*) присутствует последовательность PBS (primer-binding site - сайт связывания праймера). Между последней (*pol* или добавочной) ОРС и 3' LTR располагается полипуриновый тракт (PPT - polypurine track) [10,91,93–95].

Процесс обратной транскрипции начинается с посадки молекулы тРНК в качестве праймера-затравки на участок PBS (Рисунок 3а). RT продлевает цепь уже в виде комплементарной ДНК (кДНК), образующей минус-цепь (Рисунок 3б), синтезированную до повторяющегося участка R. Вслед за ней RNH гидролизует РНК в образовавшемся комплексе мРНК:кДНК (Рисунок 3с). Затем происходит первая смена матрицы, в ходе которой получившаяся цепь переносится на 3' конец мРНК, связывая R-участок кДНК с комплементарным ему R 3' конца мРНК (Рисунок 3д). Цепь продолжает удлиняться, проходя через PPT, внутренние ОРС вплоть до самого домена PBS, после которого цепь мРНК уже была расщеплена доменом RNH. По мере удлинения минус-цепи кДНК, RNH продолжает расщеплять матрицу мРНК, кроме участка PPT. Этот

нерасщеплённый участок становится затравкой для плюс-цепи кДНК, продлевая её до области, комплементарной U5 минус-цепи. Далее следует расщепление тРНК-затравки доменом RNH (Рисунок 3е-ф). После происходит вторая смена матрицы, где плюс-цепь кДНК перебрасывается на участок PBS минус-цепи кДНК, комплементарный ей. Минус-цепь продолжается по матрице плюс-цепи до области U3, не имевшейся на исходной мРНК. Затем плюс-цепь, перескакивая на только что завершившийся антисмысловой участок домена PBS, нарабатывается по матрице минус-цепи, возвращаясь снова к участку U5. Минус-цепь синтезируется до U3 новой матрицы плюс-цепи, и в результате получается фрагмент двухцепочечной ДНК (дцДНК) ретротранспозона с участками U3-R-U5 с каждого конца (Рисунок 3g-h) [91,93,95]. Поскольку в ходе ретротранспозиции 3' последовательность LTR формируется с комплементарного 5' LTR участка минус-цепи, идентичность последовательностей LTR может считаться одним из ключевых маркеров недавней активности LTR-ретротранспозона.

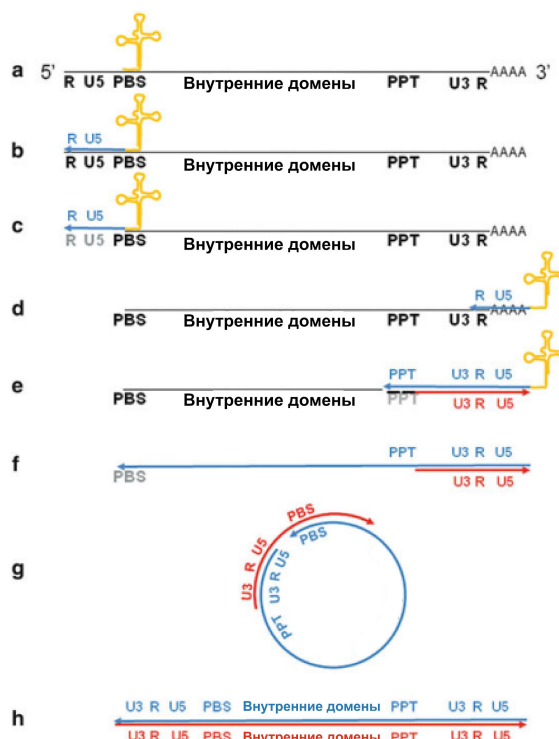


Рисунок 3. Механизм обратной транскрипции LTR-ретротранспозонов. а-h - стадии процесса обратной транскрипции. R - повторяющаяся последовательность, U5 и U3 - уникальные области на 5' и 3' конца элемента, PBS - сайт связывания праймера, PPT полипуриновый тракт. Адаптировано из [95].

Далее, готовый фрагмент ДНК интегрируется в новый локус генома за счёт активности домена INT. Домен INT, произошедший от транспозаз ДНК-

транспозонов [96–100], работает аналогичным образом [101]: (1) захватывает встраиваемый фрагмент в виде петли с концевыми повторами (LTR, IR), формирующих шпильку; (2) повторенные последовательности расщепляются с одного конца каждая; (3) комплекс садится на специфично распознаваемый интегразой сайт; (4) производит по нему разрыв цепи, в ходе которого укороченная цепь ДНК вируса/ретроэлемента сшивается с хромосомной; (5) комплекс разбирается. Далее, следует амплификация недостающих концов, удаление избыточного фрагмента 5` краёв интегрируемой ДНК и лигирование, в результате которых имеется новый локус встройки [101,102]. Схема вышеописанных событий отражена на примере ретровирусной интегразы на рисунке 4.

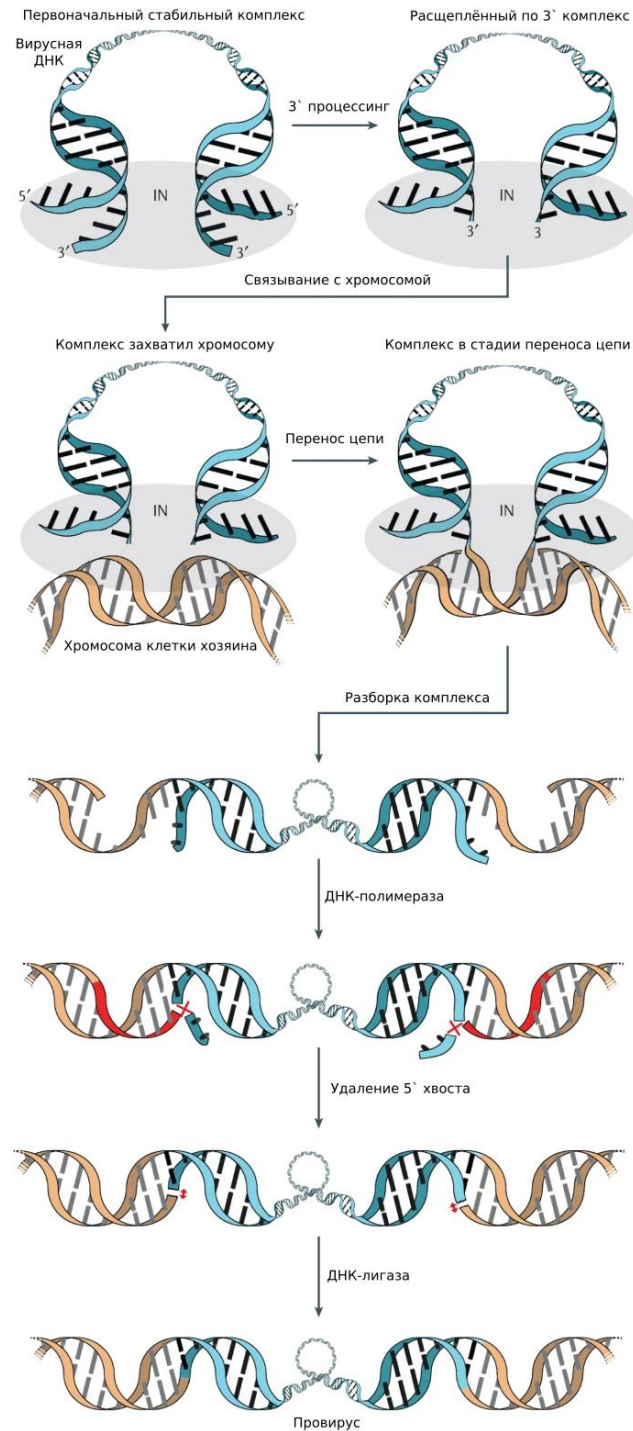


Рисунок 4. Механизм интеграции вирусной ДНК в хромосому генома-хозяина. Адаптировано из [102].

Однако, домен INT не является единственным способом интеграции ДНК ретровирусов и LTR-ретротранспозона в геном хозяина [103–106]. Так, менее эффективный (примерно на 2-4 порядка) INT-независимый путь интеграции, способный функционировать у ретровирусов и LTR-ретротранспозонов параллельно с INT-опосредованным, реализуется через путь одноцепочечного отжига (single-strand annealing (SSA) pathway). Этот путь основан на клеточных механизмах репарации двухцепочечных разрывов ДНК организма хозяина, в

частности, гомологичной рекомбинации [103,104,107]. Некоторые компоненты (mre11, rad50, rad51 [103], и в особенности ключевой элемент процесса - rad52 [103,108,109]) рекомбинационного комплекса показали критическую роль в осуществлении INT-независимой рекомбинации. Свидетельство независимой от наличия домена интегразы пути внедрения собственной ДНК отражает приспособительную роль приобретения домена INT ретровирусами и LTR-ретротранспозонами, изначально способных обходиться без него.

Функция домена RNH в LTR-ретротранспозонах

Рибонуклеаза H известна как самостоятельный фермент, относящийся к эндонуклеазам, который присутствует практически у всех живых организмов. Как у всех эндонуклеаз, основной его функцией является расщепление нуклеиновых кислот по фосфодиэфирной связи [110,111]. Существует три типа фермента RNH, схожие по структуре, но различающиеся по предпочитаемому субстрату. RNH, свойственная LTR-ретротранспозонам, относится к первому типу (RNase HI, RNH 1 type), который выявлен у представителей про- и эукариот и гидролизует РНК преимущественно в комплексах РНК с ДНК [110].

Общая часть структуры фермента состоит из 5 β -складок, формирующих β -лист, и 3 α -спиралей (обычно их больше, но не все участвуют в формировании активного центра) [112–114]. На них располагаются аминокислотные остатки каталитически активного центра - тетрада аминокислот DEDD. Однако, в тетраде DEDD только первые три аминокислоты являются необходимыми. Последний аминокислотный остаток аспарагиновой кислоты мотива DEDD может теряться у некоторых групп организмов с частичным сохранением свойств фермента. Ферменту для работы необходимо в качестве кофакторов молекула воды и два иона двухвалентных металлов. Обычно данным ионом является Mg^{2+} , реже Mn^{2+} . В результате гидролиза РНК на рибозных остатках остаются гидроксильная (-OH) группа на 3' конце и фосфорный остаток на 5' конце [110,115,116]. Реакция гидролиза по фосфодиэфирной связи представлена в общем виде на рисунке

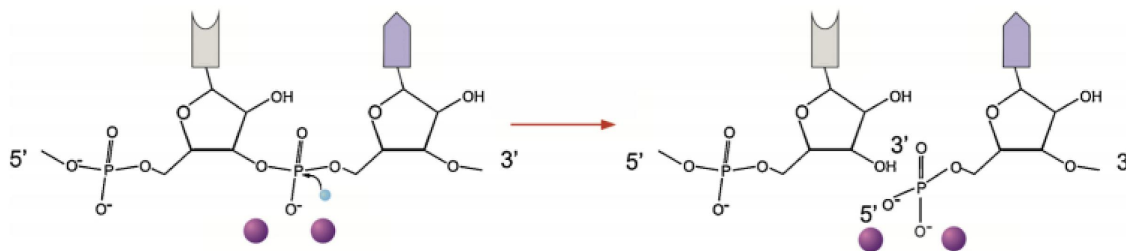


Рисунок 5. Реакция гидролиза нуклеиновой кислоты ферментом RNH. Фиолетовым цветом обозначены атомы Mg^{2+} , синим - молекула воды, нуклеофильно атакующая фосфор. Адаптировано из [110].

В свободной форме фермент RNH типа 1 самостоятельно “выгибает” молекулу субстрата, чтобы конформационные изменения способствовали гидролизу [117]. В LTR-ретротранспозонах RNH не является самостоятельным ферментом, а состоит в мультидоменном ферменте вместе с RT. При этом, в силу деградации домена RNH (отсутствие факультативной пятой аминокислоты активного центра - аргинина или гистидина, иногда участвующей в его работе, потеря четвёртой аминокислоты тетрады некоторыми кластерами ретротранспозонов, общая деградация С-концевого участка домена), функцию связывания субстрата перенимает домен RT. Конформационные изменения домена RT влекут за собой изменение активности связанного с ней домена RNH, чем, вероятно, достигается специфичность взаимодействия RNH ретротранспозонов с комплексом РНК:кДНК. Также предполагается, что именно благодаря этой деградации домена RNH LTR-ретротранспозонов фермент не разрушает область PPT, способствуя образованию плюс-цепи ДНК LTR-ретротранспозона [94,110].

Отдельные группы LTR-ретротранспозонов независимо приобрели в процессе своей эволюции дополнительный домен RNH: группа *Tat* зеленых растений [9] и группы *Chronos* и *Archon* оомицетов [118]. Заново приобретённый домен RNH не имеет столь же выраженной деградации собственной последовательности, а по своему происхождению оказался филогенетически более близким к генам RNH 1 типа бактерий, архей и растений. Названный “архейным” RNH (Archaea-like RNH, aRNH), “новый” домен был выявлен в нескольких позициях внутри гена *pol* [9,118].

***Tat* LTR-ретротранспозоны**

Кластер *Tat* из LTR-ретротранспозонов подкласса *Ty3/gypsy* был обнаружен только в геномах зеленых растений и является одной из наиболее сложно устроенных групп LTR-ретротранспозонов. Их длина составляет от 10 до 21 т.п.о., повторы LTR имеют длину до 1000 п.о. Помимо общего для всех LTR-ретротранспозонов строения элементы *Tat* цветковых растений могут содержать дополнительные ОРС с неизвестными функциями [119]. Обычно эти ОРС располагаются до гена *gag* и после гена *pol* в смысловом и антисмысловом положениях [7].

Наиболее интересной особенностью элементов кластера *Tat* является наличие закодированного в гене *pol* дополнительного домена aRNH [9]. На рисунке 6 представлена в общем виде структура *Tat* LTR-ретротранспозонов.

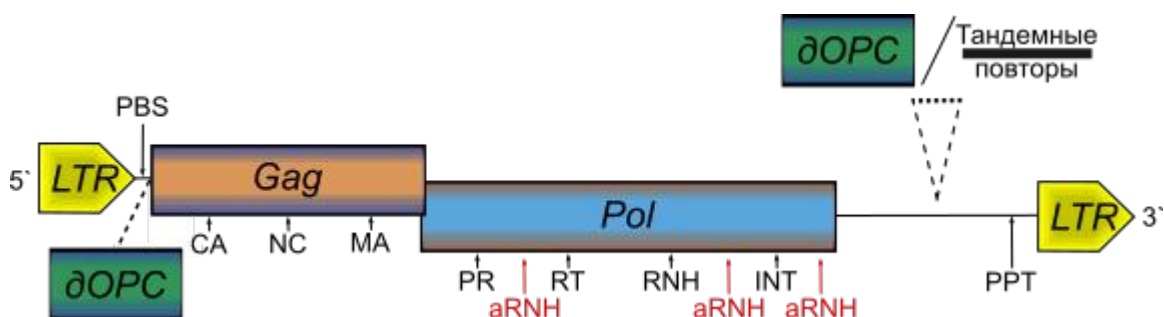


Рисунок 6. Обобщенная структура *Tat* LTR-ретротранспозонов. Gag и Pol - две основные ОРС. CA - капсид, NC - нуклеокапсид, MA - матричный домен (matrix), PR - аспартильная протеаза, RT - обратная транскриптаза, RNH - рибонуклеаза H, INT - интеграза, LTR - длинные концевые последовательности, ΔОРС - дополнительные открытые рамки считывания, PPT - полипуриновый тракт, PBS - сайт связывания праймера, aRNH - возможные положения дополнительного домена RNH, выявленные в прежних работах [9].

Стоит обратить внимание, что на иллюстрации отражено всё известное на данный момент многообразие вариантов положения домена aRNH в гене *pol*. Это непостоянство объясняется тем, что среди элементов *Tat* выявлены различные структурные варианты, содержащие домен в различных положениях. Структурные варианты, выявленные в прежних работах лаборатории молекулярно-генетических систем ИЦиГ СО РАН представлены на рисунке 7.

Распределение структурных вариантов по таксонам зеленых растений также представляет глубокий интерес. В прежних работах поиск *Tat* LTR-ретротранспозонов выполнялся преимущественно на геномах покрытосеменных

(цветковых) растений [9]. Все элементы *Tat* LTR-ретротранспозонов, выявленные в этой группе растений, имели единый структурный вариант по гену *pol*, обозначенный на рисунке 7 “*Tat* IV-VI”. У других изученных в той же работе геномов (*Selaginella moellendorffii*; класс Lycopodiopsida и три генома хвойных - класс Pinopsida) наблюдались два других структурных варианта по положению *aRNH* (группы I, II и III).

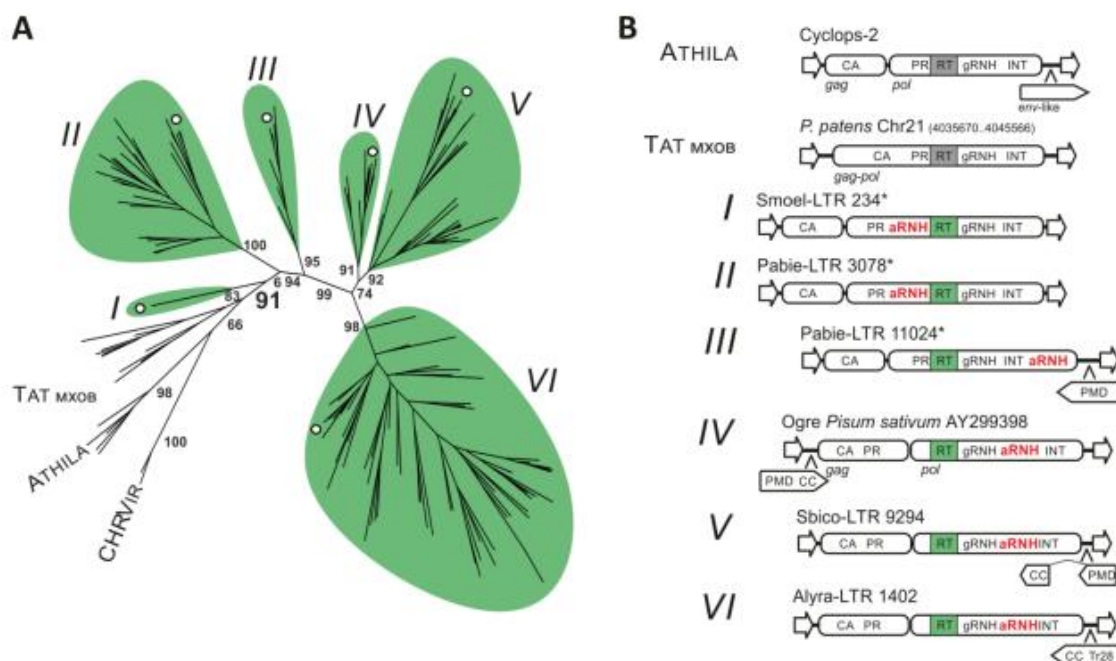


Рисунок 7. А. Филогенетическое дерево кластера *Tat/Athila* группы *Ty3/gypsy* LTR-ретротранспозонов. Ветвь *CHRVir* (хромовирусы) является внешней группой на данном дереве. В. Структурные варианты *Tat* LTR-ретротранспозонов, выявленные у высших растений и их ближайшая родственная группа - кластер *Athila*. Римскими цифрами пронумерованы кластеры элементов *Tat* LTR-ретротранспозонов из работы [9]. CA - капсид, PR - аспартильная протеаза, RT - обратная транскриптаза, gRNH - исходный домен рибонуклеазы Н, INT - интегразы, aRNH - дополнительный, “архейный”, домен рибонуклеазы Н, env-like - домен подобный домену в гене оболочки ретровирусов (env). CC, PMD, Tr28 - консервативные белковые домены с неизвестной функцией в составе дополнительных ОРС. Адаптировано из [9].

Конвергентная эволюция групп ретротранспозонов, имеющих домен *aRNH*

Тенденция приобретения дополнительного домена *aRNH* была выявлена не только у элементов *Tat* LTR-ретротранспозонов. До *Tat* LTR-ретротранспозонов было известно, что ретровирусы позвоночных также приобрели второй ген *RNH* [115]. Считается, что новый ген *RNH* был приобретён в ходе эволюции ретровирусов позвоночных из генома организма-

хозяина, в то время как исходный ген RNH присутствовал ещё у предка ретровирусов - древнего *Ty3/gypsy* LTR-ретротранспозона [120,121]. Впоследствии в этой “паре” доменов RNH произошла специализация, в ходе которой оригинальная копия RNH преобразовалась в так называемый домен-”связку” (tether domain), который потерял каталитическую активность [8,115,122]. Считается, что за счёт данной “связки” регулируется активность нового домена RNH, повышается активность переноса цепи ДНК в ходе обратной транскрипции и повышается частота рекомбинации в молекулах вирусной РНК, ускоряя эволюцию вируса в гонке вооружений с иммунной системой организма-хозяина [115].

Пара доменов RNH-aRNH у *Tat* LTR-ретротранспозонов цветковых растений и пара “связка”-RNH у ретровирусов позвоночных демонстрируют одинаковую тенденцию как приобретения ретроэлементом последовательности гена RNH 1 типа из генома хозяина, так и взаимодействие нового домена с исходным доменом RNH и RT [9]. В каждой из групп сохраняется общая структура, где новый домен RNH следует за исходным. Таким образом, образуется схожий порядок доменов “RT-RNH-aRNH” у *Tat* LTR-ретротранспозонов и “RT-tether-RNH” ретровирусов позвоночных. Повторное приобретение гена RNH от разных хозяев и общность их положения свидетельствует о конвергентной эволюции ретровирусов позвоночных и LTR-ретротранспозонов растений к данному структурному “фенотипу”. Вероятно, данное положение нового домена RNH у ретровирусов и aRNH у *Tat* LTR-ретротранспозонов поддерживается отбором [9]. Перечисленные события проиллюстрированы на рисунке 8.

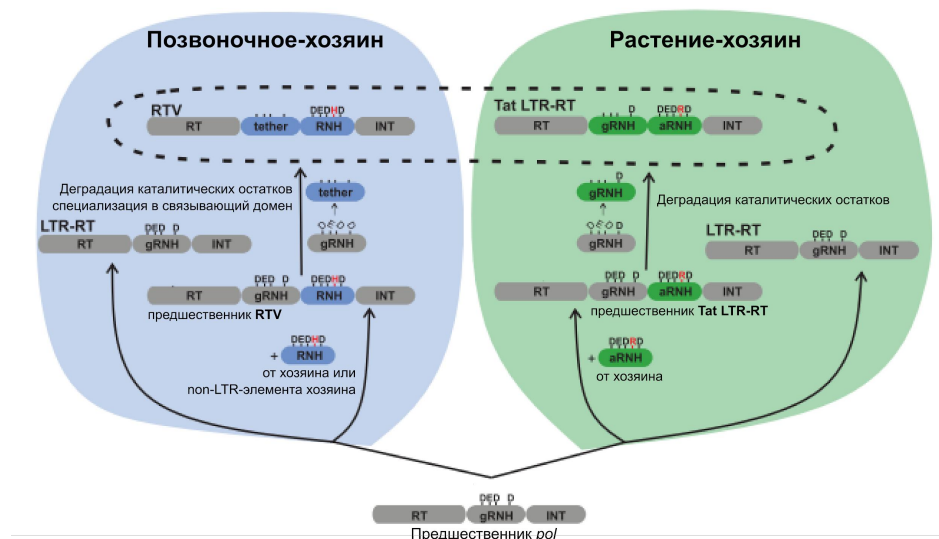


Рисунок 8. Происхождение *Tat* LTR-ретротранспозонов и ретровирусов позвоночных и их конвергентной доменной структуры (“RT-RNH-aRNH” и “RT-tether-RNH” соответственно) от общего предка *Ty3/gypsy* LTR-ретротранспозона. RVT - ретровирусы, LTR-RT - LTR-ретротранспозоны, RT - обратная транскриптаза, gRNH - нативная рибонуклеаза H, INT - интеграза, aRNH - “архейная” рибонуклеаза H tether - домен-“связка”. Над всеми структурами, производными от древнего гена RNH подписаны сохранившиеся в них аминокислоты из активного центра (мотив DEDD) и добавочная пятая аминокислота (помечена красным). Адаптировано из [9].

Также приобретение домена aRNH было выявлено у LTR-ретротранспозонов в геномах паразитических эукариот группы оомицетов (*Stramenopiles*, *Oomycetes*) [118]. Ретроэлементы оомицетов, которые приобрели домен aRNH, принадлежат к двум кластерам группы *Ty3/gypsy* LTR-ретротранспозонов: *Chronos* и *Archon*. Оба этих кластера, согласно филогенетическому анализу последовательности RT-домена, генетически удалены от ретроэлементов кластера *Tat*. Приобретённый домен aRNH в каждой из этих групп следует сразу за доменами RT и исходным доменом RNH. Таким образом, ещё в одной группе LTR-ретротранспозонов, помимо ретровирусов позвоночных животных и *Tat* LTR-ретротранспозонов цветковых растений, наблюдается один и тот же структурный вариант гена *pol*. Что позволяет в очередной раз утверждать, что возникая независимо, данный порядок доменов поддерживается отбором, и, следовательно, благоприятен для жизнедеятельности LTR-ретротранспозонов [118]. Строение LTR-ретротранспозонов оомицетов из групп *Chronos* и *Archon* представлено на

рисунке

9.

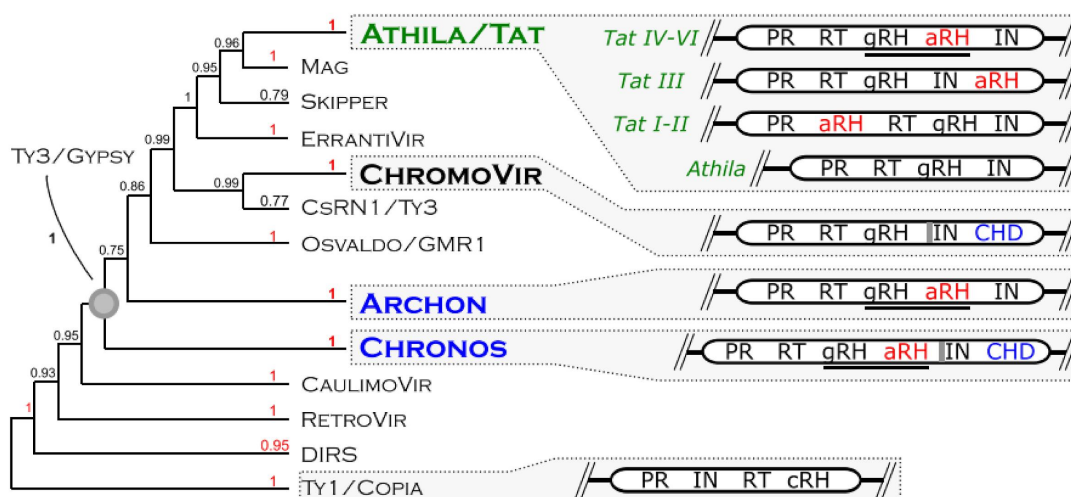


Рисунок 9. Филогенетическое древо LTR-ретротранспозонов, построенное по домену RT. Зелёным цветом выделены группы и кластеры, обнаруженные у высших растений, синим - у оомицетов. На узлах дерева подписаны значения статистической поддержки aLRT (approximate likelihood-ratio test), красным выделены узлы, с дополнительной бутстреп-поддержкой более 60 из 100 итераций. PR - аспартильная протеаза, RT - обратная транскриптаза, gRH - нативная рибонуклеаза H, INT - интеграз, aRH - “архейная” рибонуклеаза H (выделены красным цветом в структурах консенсусных элементов). Адаптировано из [118].

Заключение

Тенденция разных групп LTR-ретротранспозонов независимо друг от друга приобретать дополнительный домен RNH, а также закрепление нового домена в виде “ретровирусной” структуры “RT-RNH-aRNH” (или “RT-tether-RNH”) показана на трёх разных группах LTR-ретроэлементов из геномов зеленых растений (*Tat*), паразитических протист оомицетов (*Chronos* и *Archon*) и предшественников ретровирусов позвоночных животных. Такая тенденция, по всей видимости, поддерживается отбором и несет важное приспособительное значение для жизненного цикла ретротранспозонов именно со структурой “RT-RNH-aRNH”. Однако эволюционная роль данного приспособления все еще недостаточно изучена.

В группе LTR-ретротранспозонов *Tat* у нецветковых растений выявлены также другие структурные варианты расположения приобретённого домена aRNH (Рисунок 7). Механизм закрепления именно структурного варианта “RT-RNH-aRNH” остаётся неясным. Изучение разнообразия структурных вариантов

Tat может пролить свет на эволюционные процессы, имевшие место в разных группах LTR-ретротранспозонов от момента приобретения дополнительного домена RNH до закрепления конкретного структурного варианта.

Глава 2. Методы исследования

Прежний алгоритм поиска и извлечения последовательностей LTR-ретротранспозонов был основан на первоначальном поиске последовательностей LTR. Для этой задачи использовалось программное обеспечение LTRharvest [123], которое обнаруживает повторы заданной длины и идентичности на заданном расстоянии. Затем, внутри обнаруженной пары LTR ищались домены RT и aRNH [9]. Однако копии ретротранспозонов, которые длительное время пребывают в геноме без перемещения, неизбежно накапливают мутации в последовательностях LTR [124]. Большое число таких мутаций приводит последовательности LTR к деградировавшему состоянию, и чувствительность программы не позволяет их обнаружить. Ввиду вероятной (высокой) недооценки количества искомым ретротранспозонов было принято решение разработать новый алгоритм поиска и извлечения последовательностей LTR-ретротранспозонов, основанный на первичном поиске консервативных белковых последовательностей доменов RT и aRNH (подробное описание работы алгоритма далее).

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>). В базе данных plaBi представлен список растений, для которых имеется прочтённый геном и ссылка на базу данных с ним. Список полученных геномных сборок и систематическое положение видов представлены в таблице 1.

Алгоритм поиска и извлечения *Tat* LTR-ретротранспозонов

Автоматизированный алгоритм поиска, извлечения последовательностей и анализа структур *Tat* LTR-ретротранспозонов в исследуемых геномах растений написан на языке программирования Python (версия 3.6) с применением пакета Biopython [125]. Сокращённая схема работы алгоритма изображена на рисунке 10.

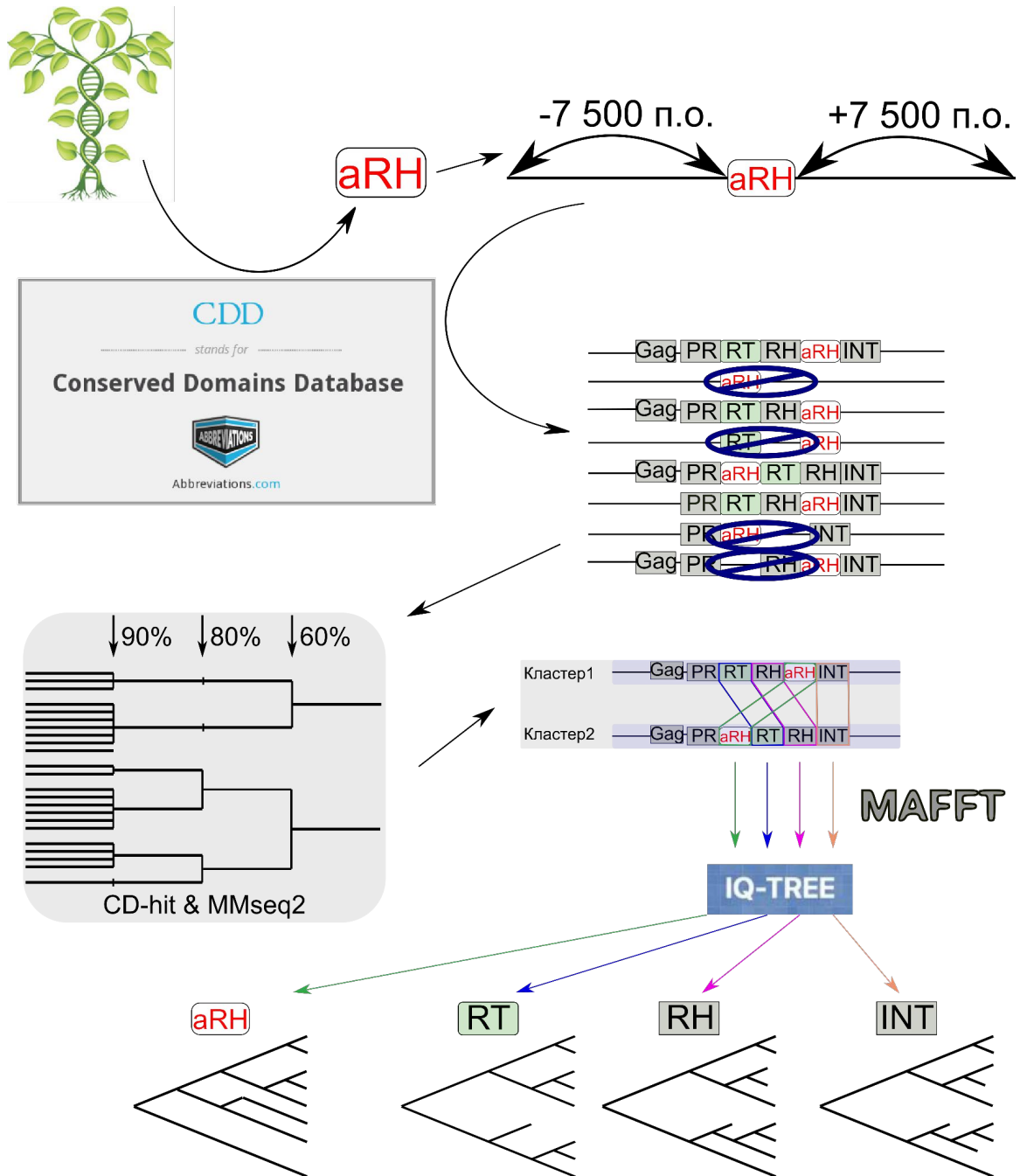


Рисунок 10. Сокращённая схема работы описываемого алгоритма. Файлы геномной сборки подаются на инструмент RPS-BLAST, работающий с базой данных CDD, на первом шаге ограниченной только 1 доменом для поиска. В данном случае это домен aRNH (обозначен как aRH). Затем, по координатам для каждого домена извлекается его фланкирующая область (координаты ± 7500 пар оснований), которая подаётся на второй раунд поиска доменов. По координатам восстанавливается доменный состав вероятного элемента, отфильтровываются непригодные для анализа последовательности. Оставшиеся элементы кластеризуются по последовательности центрального домена RT инструментами CD-hit или MMseqs2. Выбираются элементы-представители полученных кластеров. От них берутся 4 домена (RT, INT, RNH и aRNH, обозначенные как RH и aRH соответственно), три из которых основные в LTR-ретротранспозонах, а последний, aRNH, ключевой в данном исследовании. Для каждого домена строится выравнивание инструментом MAFFT и впоследствии производится филогенетическая реконструкция инструментом IQTREE.

Алгоритм осуществляет: (1) биоинформатический поиск гомологичных белковых последовательностей с помощью программного пакета RPS-BLAST [126] в нуклеотидных геномных последовательностях; (2) анализ, извлечение и классификацию структурных доменов LTR-ретротранспозонов на основании данных RPS-BLAST и локальной версии программы для определения белковых доменов CD-search [127]; (3) реконструкцию структурных вариантов *Tat* LTR-ретротранспозонов.

На данный момент алгоритм имеет несколько версий, различающихся условиями поиска и возможностью настраивать их при запуске. Пошаговое описание принципов работы алгоритма представлено ниже.

Шаг 1. Первоначально был разработан алгоритм, выполняющий поиск инструментом RPS-BLAST последовательностей домена aRNH среди нуклеотидных последовательностей геномных сборок. Впоследствии этот блок был переработан для первоначального поиска домена RT. На данный момент имеется возможность выбора домена для первого шага (RT или aRNH). 94 исследованных генома прошли через каждый из двух путей данного поиска. Поиск через домен RT значительно расширяет выборку полученных последовательностей, позволяя выявлять не только элементы кластера *Tat* (т.е. содержащие домен aRNH), но и элементов остальных кластеров *Ty3/gypsy* LTR-ретротранспозонов. Это указывает на перспективность и масштабируемость алгоритма для работы с другими группами LTR-ретротранспозонов.

На данном этапе алгоритм также был специально адаптирован для поиска доменов RT других групп LTR-ретротранспозонов (*Ty1/copia*, *Ty3/gypsy* и *Bel/Pao*, *DIRS*, *PLEs* - *Penelope*-подобные элементы) и ретровирусов. В ранней версии (производившей поиск элементов *Ty1/copia*, *Ty3/gypsy* и *Bel/Pao*) алгоритм был успешно апробирован для аннотации LTR-ретротранспозонов в геноме свободноживущего регенерирующего модельного плоского червя *Macrostomum lignano*.

Неожиданным оказалось то, что поиск через домен RT позволил расширить выборку элементов с aRNH, в которых домен aRNH располагался в отдельной OPC (5' или 3', отдельно от OPC генов *gag* и *pol*). В более поздней версии алгоритма в этой части скрипта были внесены правки, позволяющие поиску через домен aRNH не терять недостающее количество элементов.

По окончании поискового этапа последовательности искомого домена вместе с фланкирующими его областями из геномной сборки извлекаются скриптом с использованием пакета Biopython. Суммарная длина фланкирующих областей и последовательности между ними была сопоставима со средней длиной *Tat* LTR-ретротранспозонов. Данный блок “извлечения” един для всех последующих версий скрипта.

Шаг 2 алгоритма начинается с повторного поиска доменов инструментом RPS-BLAST. Однако на данном шаге поиск производится против локальной базы данных, составленной из основных доменов LTR-ретротранспозонов (в частности, домена GAG из соответствующего гена *gag* и доменов PR, RT, RNH, INT гена *pol*). Для всех обнаруженных доменов (включая aRNH) извлекается аминокислотная последовательность, информация о ОПС и координаты доменов, расположенных по краям на случай, если не удастся обнаружить последовательности LTR. На основании наборов координат найденных доменов алгоритм также определяет и выводит в названии последовательности элемента его структуру по доменам. Поиск пары последовательностей LTR производится инструментом BLASTn из пакета BLAST+ [126]. На вход BLASTn подаются области, фланкирующие начало первого и конец последнего найденного домена. Рассматривались ближайшие к доменам пары последовательностей с высокой идентичностью (более 80%) и длиной не менее 100 нуклеотидов. По полученным координатам краёв последовательностей LTR формировалась окончательная нуклеотидная последовательность элемента.

Для дальнейшего анализа производилась фильтрация элементов по наличию трёх ключевых доменов для жизнедеятельности LTR-ретротранспозонов - RT, RNH, INT. Данные домены также обладают наиболее консервативными аминокислотными последовательностями, благодаря чему возможно их использование для построения достоверной филогении ретротранспозонов из эволюционно далеких видов.

Последним процессом второго шага является кластеризация полученных элементов и выбор элементов-представителей от кластеров. В силу особенностей эволюции ретротранспозонов наиболее подходящим кандидатом для кластеризации является домен RT. Он является центральным доменом в эволюции данной группы мобильных элементов и уже использовался прежде для описания эволюции всего класса ретротранспозонов [4]. Кластеризация сначала

производилась инструментом CD-hit [128]. Ранняя версия алгоритма с применением данного программного обеспечения (ПО) была запущена и частично проработана на 49 геномах за период первого года с момента смены темы. Однако, его алгоритм не соответствовал потребностям данной работы, потому программа была заменена ПО MMseqs2 [129], имеющим несколько различных режимов для кластеризации (easy-linclust, easy-cluster и т.д.). Выбор ПО для кластеризации, а также режима кластеризации (идентичность 50% и 80% для MMseqs2, 63% и 80% для CD-hit) доступны для выбора в поздних версиях скрипта.

Шаг 3. Реконструкция эволюционных взаимоотношений найденных структурных вариантов *Tat* LTR-ретротранспозонов производилась за два раунда построения филогенетических деревьев.

Сначала последовательности каждого из 4 доменов (RT, RNH, INT и aRNH), полученных от элементов-представителей, выравнивались инструментом для множественного выравнивания Mafft [130,131] против соответствующих доменов из базы данных GyDB [132].

Далее, начинался первый раунд, в ходе которого инструментом FastTree [133] строилось “быстрое” филогенетическое древо для каждого из 4 доменов. Данный инструмент позволяет получить примерные данные о филогенетическом взаимоотношении исследуемых последовательностей. Также FastTree позволяет предварительно оценить: насколько у выбранных элементов-представителей различается эволюционная судьба их отдельно взятых белковых доменов. Для группы кластеров, сформировавших общую ветвь (кластер) на древе FastTree был выбран кластер-представитель, используемый на следующем раунде построения филогенетического древа. Затем для основных доменов - RT, RNH, INT - алгоритм производил конкатенацию аминокислотных последовательностей в перечисленном порядке. Конкатенация производилась с выравниванием данных доменов отдельно написанным скриптом на языке программирования Python. Полученные “конкатенаты” использовались для построения более точного филогенетического древа (см. далее).

Второй раунд построения деревьев для конкатенированных последовательностей трёх доменов и отдельно для домена aRNH производился инструментом IQTree [134,135], гарантирующим большую точность представленных на древе взаимоотношений элементов. Статистическая

поддержка достоверности полученных узлов (кластеров) вычислялась алгоритмами “приблизительного” теста правдоподобия (approximate likelihood-ratio test, aLRT [136]) и ультра-быстрого бутстрепа (ultra-fast bootstrap, ufBoot [137]). Достоверными принимались узлы, значение которых было выше 80 и 95 для aLRT и ufBoot, соответственно.

Таблица 1. Систематическая классификация по таксонам “отдел”, “класс” и “вид” растений, имеющих прочтённые геномы, и которые были исследованы в данной работе. Также представлены базы данных, из которых получены соответствующие геномные сборки. Представлены базы данных: NCBI:Genome (<https://www.ncbi.nlm.nih.gov/genome>), GigaDB (<http://gigadb.org/site/index>), FernBase (<https://www.fernbase.org/>), Phytozome (<https://phytozome.jgi.doe.gov/pz/portal.html>), DRYAD (<https://datadryad.org/>), CNGBdb (<https://www.cngb.org/index.html>), NGDC (<https://bigd.big.ac.cn/>), JGI (<https://genome.jgi.doe.gov/portal/>), TreeGenesDB (<https://treegenesdb.org/>). Использован также сайт института Цюриха (<https://www.uzh.ch/en.html>).

No	Основные группы	Класс	Вид	База данных
1	Тип Chlorophyta (Зелёные водоросли)	Palmophyllophyceae	<i>Prasinoderma_coloniale</i>	CNGBdb
2		Mamiellophyceae	<i>Ostreococcus lucimarinus</i>	NCBI:Genome
3			<i>Ostreococcus tauri</i>	NCBI:Genome
4			<i>Bathycoccus prasinos</i>	NCBI:Genome
5			<i>Micromonas pusilla</i>	NCBI:Genome
6			<i>Micromonas commoda</i>	NCBI:Genome
7		Pyramimonadophyceae	<i>Cymbomonas tetramitiformis</i>	NCBI:Genome
8		Chlorophyceae	<i>Chlamydomonas reinhardtii</i>	NCBI:Genome
9			<i>Chlamydomonas eustigma</i>	NCBI:Genome
10			<i>Volvox carteri</i>	NCBI:Genome
11			<i>Gonium pectorale</i>	NCBI:Genome
12			<i>Tetrabaena socialis</i>	NCBI:Genome
13			<i>Haematococcus pluvialis</i>	NCBI:Genome
14			<i>Yamagishiella unicocca</i>	NCBI:Genome
15			<i>Eudorina sp.</i>	NCBI:Genome
16			<i>Mychonastes homosphaera</i>	NCBI:Genome
17			<i>Monoraphidium neglectum</i>	NCBI:Genome
18			<i>Chromochloris zofingiensis</i>	NCBI:Genome
19			<i>Tetradismus obliquus</i>	NCBI:Genome

20			<i>Raphidocelis subcapitata</i>	NCBI:Genome
21			<i>Scenedesmus quadricauda</i>	NCBI:Genome
22		Ulvophyceae	<i>Ulva mutabilis</i>	NCBI:Genome
23			<i>Caulerpa lentillifera</i>	NCBI:Genome
24		Trebouxiophyceae	<i>Parachlorella kessleri</i>	NCBI:Genome
25			<i>Chlorella variabilis</i>	NCBI:Genome
26			<i>Chlorella sorokiniana</i>	NCBI:Genome
27			<i>Micractinium conductrix</i>	NCBI:Genome
28			<i>Auxenochlorella protothecoides</i>	NCBI:Genome
29			<i>Auxenochlorella pyrenoidosa</i>	NCBI:Genome
30			<i>Helicosporidium sp.</i>	NCBI:Genome
31			<i>Prototheca zopfii</i>	NCBI:Genome
32			<i>Chloroidium sp</i>	NCBI:Genome
33			<i>Coccomyxa subellipsoidea</i>	NCBI:Genome
34			<i>Picochlorum sp.</i>	NCBI:Genome
35			<i>Picochlorum costavermella</i>	NCBI:Genome
36			<i>Botryococcus braunii</i>	NCBI:Genome
37		Klebsormidiophyceae	<i>Klebsormidium flaccidum</i>	NCBI:Genome
38		Charophyceae	<i>Chara braunii</i>	NCBI:Genome
39	Тип Streptophyta	Chlorokybophyceae	<i>Chlorokybus atmophyticus</i>	NCBI:Genome
40	(“Стрептофитовые” водоросли, не включая высших растений)	Mesostigmatophyceae	<i>Mesostigma viride</i>	NCBI:Genome
41		Zygnemophyceae	<i>Mesotaenium endlicherianum</i>	NCBI:Genome
42			<i>Spirogloea muscicola</i>	NCBI:Genome
43	Клада Marchantiophyta	Marchantiopsida	<i>Marchantia polymorpha</i>	NCBI:Genome
44	(Печёночные мхи)		<i>Marchantia inflexa</i>	NCBI:Genome
45	Клада Anthocerotophyta	Anthocerotopsida	<i>Anthoceros angustus</i>	NCBI:Genome
46	(Антоцеротовые мхи)		<i>Anthoceros punctatus</i>	uzh.ch
47			<i>Anthoceros agrestis</i>	uzh.ch
48	Клада Bryophyta	Sphagnopsida	<i>Sphagnum magellanicum</i>	Phytozome
49	(Настоящие мхи)		<i>Sphagnum fallax</i>	Phytozome
50		Bryopsida	<i>Physcomitrella patens</i>	NCBI:Genome

51			<i>Pleurozium schreberi</i>	NCBI:Genome
52			<i>Calohypnum plumiforme</i>	GSA
53			<i>Ceratodon purpureus</i>	Phytozome
54			<i>Fontinalis antipyretica</i>	CNGBdb
55			<i>Syntrichia caninervis</i>	NCBI:Genome
56	Класс Lycopodiopsida (плауны)	Lycopodiopsida	<i>Selaginella tamariscina</i>	NCBI:Genome
57			<i>Selaginella moellendorffii</i>	NCBI:Genome
58			<i>Selaginella kraussiana</i>	NCBI:Genome
59			<i>Isoetes engelmannii</i>	NCBI:Genome
60	Класс Polypodiopsida (Папоротникообразные)	Polypodiopsida	<i>Polypodium sp</i>	https://digitalcommons.usu.edu/fern_genome/2
61			<i>Pteridium sp</i>	
62			<i>Plagiogyria sp</i>	
63			<i>Dipteris sp</i>	https://digitalcommons.usu.edu/fern_genome/2
64			<i>Cystopteris sp</i>	
65			<i>Ceratopteris sp</i>	
66			<i>Ceratopteris richardii</i>	NCBI:Genome
67			<i>Salvinia cucullata</i>	Fernbase
68			<i>Azolla filiculoides</i>	Fernbase
69	Клада Gnetophyta (Гнетовидные)	Gnetopsida	<i>Gnetum montanum</i>	DRYAD
70			<i>Welwitschia mirabilis</i>	DRYAD
71	Клада Ginkgophyta (Гинкговые)	Ginkgoopsida	<i>Ginkgo biloba</i>	GigaDB
72	Клада Cycadophyta (Саговниковидные)	Cycadopsida	<i>Cycas panzhihuaensis</i>	NCBI:Genome
73	Класс Pinopsida (Голосеменные)	Pinopsida (conifers I clade - отдел Pinales, хвойные)	<i>Pseudotsuga menziesii</i>	NCBI:Genome
74			<i>Abies alba</i>	TreeGenesDB
75			<i>Pinus taeda</i>	NCBI:Genome
76			<i>Pinus lambertiana</i>	NCBI:Genome
77			<i>Picea abies</i>	NCBI:Genome
78			<i>Larix sibirica</i>	NCBI:Genome
79			<i>Picea sitchensis</i>	NCBI:Genome
80			<i>Larix kaempferi</i>	NCBI:Genome

81			<i>Picea glauca</i>	NCBI:Genome
82		Pinopsida (conifers II clade - отдел Cupressales, кипарисовые и тисовые)	<i>Sequoiadendron giganteum</i>	NCBI:Genome
83			<i>Sequoia sempervirens</i>	NCBI:Genome
84			<i>Thuja plicata</i>	Phytozome
85			<i>Taxus wallichiana</i>	NCBI:Genome
86	Класс Magnoliopsida (Покрытосеменные)	Magnoliopsida	<i>Amborella trichopoda</i>	NCBI:Genome
87			<i>Nymphaea colorata</i>	NCBI:Genome
88			<i>Nymphaea thermarum</i>	NCBI:Genome
89			<i>Aristolochia fimbriata</i>	NCBI:Genome
90			<i>Liriodendron chinense</i>	NCBI:Genome
91			<i>Chimonanthus salicifolius</i>	NCBI:Genome
92			<i>Cinnamomum kanehirae</i>	NCBI:Genome
93			<i>Persea americana</i>	NCBI:Genome
94			<i>Litsea cubeba</i>	NCBI:Genome

Определение зависимости деградации нативного домена RNH от наличия добавочного домена aRNH в различных положениях

На примере ретровирусов было показано, что после приобретения нового домена RNH, исходный (нативный) деградировал, теряя аминокислотные остатки каталитического центра, превращаясь в т.н. домен-”связку” (“tether”) [8,84,115]. Аналогичная тенденция была показана в прежней работе на элементах *Tat* цветковых растений с “ретровирусной” структурой [9]. Чтобы показать наличие подобной тенденции в других структурах *Tat*, производилось выравнивание доменов aRNH и нативных RNH по третичной структуре против имеющихся в базе данных PDB [138] структур этого семейства с помощью инструмента PROMALS3D [139]. На основании полученного выравнивания для каждого из выявленных сочетаний кластер-структура *Tat* выбирались наиболее сохранившиеся, интактные, последовательности генов RNH и aRNH, которые выносились в отдельное выравнивание и проводилась визуальная оценка сохранности каталитической тетрады DEDD.

Поиск дополнительных открытых рамок считывания

LTR-ретротранспозоны способны приобретать дополнительные ОРС (дОРС), кодирующие последовательности различных белковых структур. Функциональная роль преобладающей части данных структур остаётся неисследованной [140]. Однако приобретение и закрепление новой доменной структуры вне зависимости от её предназначения - весомое эволюционное событие в жизнедеятельности мобильных элементов, ведущее к дивергенции новых эволюционных линий. Поиск и изучение данных структур расширяет представление о многообразии и эволюции мобильных элементов.

Кластер *Tat*, а также родственный ему кластер *Athila*, состоящие исключительно из LTR-ретротранспозонов растений, известны за разнообразие дОРС. Среди доменов, свойственных данным группам, ранее были выявлены *env*-подобные домены (структурно соответствуют гену *env* ретровирусов, но, в силу наличия у растений клеточной стенки, его функция остаётся под вопросом), свойственные только представителям *Athila* [140]; дополнительные домены RNH [9,120], plant mobile domain (PDM), мотив coil-coiled (CC), домен транспозазы 28 (Tr28), свойственные только *Tat*, а также другие ОРС с неизвестной функцией [140].

В рамках данной работы поиск дОРС производился вручную на отдельных последовательностях элементов-представителей посредством онлайн-инструмента ORFfinder [141] и онлайн-ресурса CD-search [127]. От каждого кластера-представителя с финального филогенетического древа отбиралось около 25-30 элементов с наиболее сохранившейся структурой доменов и наиболее близкой по идентичности парой LTR. Нуклеотидная последовательность данных кластеров подавалась на онлайн-ресурс ORFfinder для выявления дОРС помимо ОРС генов *gag* и *pol*. Полученные белковые последовательности отправлялись на определение консервативных белковых последовательностей на онлайн-ресурсе CD-search. Достоверными структурами считались домены или их фрагменты со значением $e\text{-value} < 0.01$.

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 неоднократно дорабатывался и имеет несколько рабочих версий, включая финальную, результаты использования которой представлены далее. 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). Также последняя версия алгоритма автоматически распаковывает архивированный файл генома, по необходимости разбивает его на файлы меньшего размера, более подходящие для подачи алгоритму RPS-BLAST, а по окончании архивирует исходный файл обратно для сохранения пространства на жестком диске.

На данный момент алгоритм, который мы называли DARTS (Domain-Associated RetroTransposon Search), апробирован и опубликован в виде краткого методологического обзора [142], а его скрипты размещены в открытом доступе на GitHub (https://github.com/Mikkey-the-turtle/DARTS_v0.1.git). В статье производится сравнение эффективностей алгоритма DARTS с популярными и наиболее распространенными подходами, в частности с программным обеспечением, выполняющим поиск последовательностей LTR. Исследование показало преимущество алгоритма DARTS в оценке количества элементов по сравнению с методом из прежних работ [9,118], основанным на инструменте LTRharvest [123]. Эффективность работы алгоритма DARTS показана на примере геномов плауна *Selaginella moellendorffii* и трёх геномов покрытосеменных растений (*Arabidopsis thaliana*, *Nicotiana tabacum* и *Zea mays*). Сравнение эффективности работы алгоритма DARTS представлено на рисунке 11 на примере двух поисковых запросов по доменам RT (Рисунок 11a) и aRNH (Рисунок 11b), соответственно.

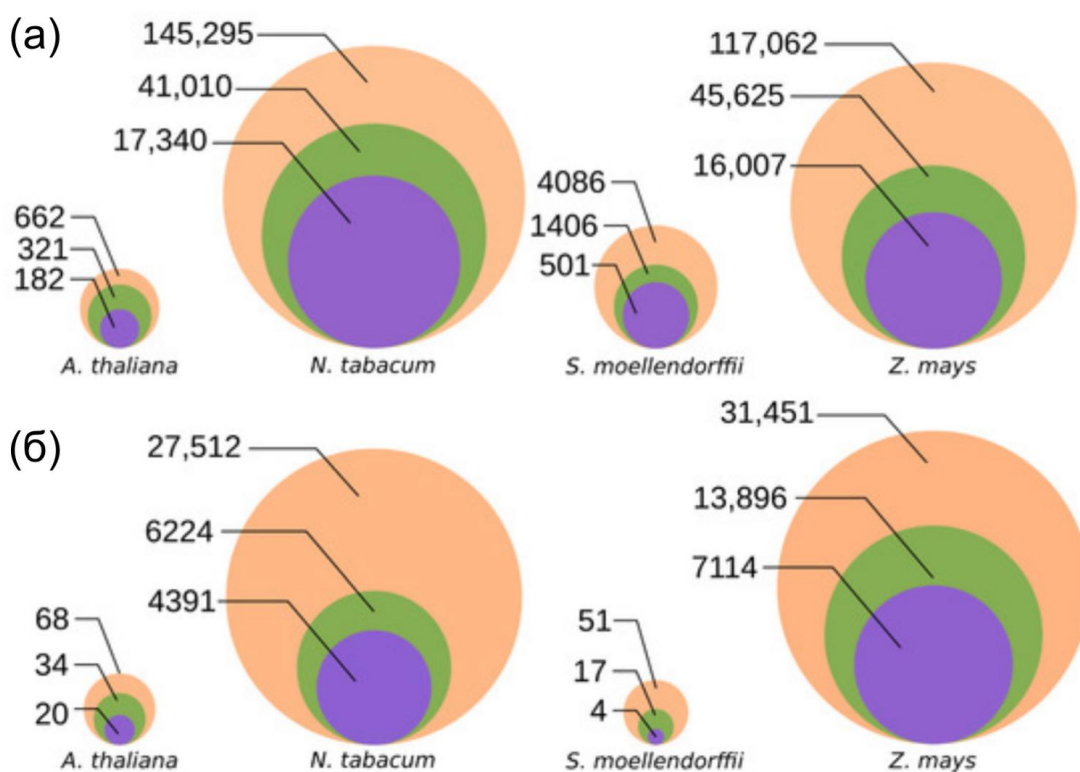


Рисунок 11. Чувствительность идентификации LTR-ретротранспозонов посредством алгоритма DARTS и алгоритма, основанного на инструменте LTRharvest. Оранжевые круги отражают количество элементов, найденных DARTS; зелёные — число элементов, найденных DARTS, содержащих последовательности LTR; фиолетовые — 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.

Используя широкие возможности к адаптации DARTS для специфичного и чувствительного поиска практически любых повторов с консервативными белок-кодирующими последовательностями, мы дополнительно апробировали алгоритм для поиска всех LTR-ретротранспозонов в геномах свободно-живущих плоских червей *M.lignano* и *Symsagittifera roscoffensis* в рамках совместной работе с лабораторией регуляции стволовых клеток и механизмов регенерации (Laboratory of Stem Cell Regulation and Mechanisms of Regeneration) Европейского исследовательского института биологии старения (European Institute for the Biology of Ageing, ERIBA, Гронинген, Нидерланды).

Потенциально, алгоритм DARTS может быть расширен для поиска любых повторенных последовательностей, содержащих белковые домены, известные базе данных CDD и взаимосвязанных с ней баз данных. Примером тому является

расширение диапазона поиска при переходе от домена aRNH к доменам группы RT, как центральным элементам при первичном поиске.

Всего в рамках данного исследования алгоритмом DARTS были проанализированы геномные сборки 94 видов зелёных растений, относящихся к разным таксономическим группам (Таблица 1).

Поиск элементов на первом шаге работы алгоритма запускали как с доменом aRNH, так и с RT. Кластеризацию выполняли с порогом идентичности 80%, снижая до порога в 50% при работе с геномами, в которых из-за высокой представленности элементов была необходимость разбивать файлы геномной сборки. Последнее пороговое значение идентичности было необходимо из-за большого количества кластеров, непосильного для адекватного процессирования инструментами множественного выравнивания.

При запуске алгоритма, начинающегося с поиска домена RT, обнаруживается приблизительно равное количество элементов по сравнению с поиском через aRNH. Незначительные различия в значениях можно объяснить наличием повреждённых или “химерных” элементов. После фильтрации и кластеризации такие элементы не были задействованы в последующем анализе, поскольку они не соответствуют выбранным критериям для элементов-представителей кластеров.

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 2, по геномам - в приложении 1.

Таблица 2. Результаты поиска LTR-ретротранспозонов методом DARTS, начинающим поиск с обнаружения домена aRNH. Представлено распределение элементов по таксонам, количеству и разнообразию структурной организации.

Таксон	Кол-во геномов	Кол-во элементов	Кол-во структур (номер структуры)
Отд. Зелёные водоросли (Chlorophyta)	36	0	0
Отд. Стрептофитовые водоросли (Streptophyta)	6	0	0
Отд. Маршанциевые мхи (Marchantiophyta)	2	0	0

Отд. Настоящие мхи (Bryophyta)	8	32	1 (#4)
Отд. Антоцеротовые мхи (Anthocerotophyta)	3	32	2 (#1,#5)
Отд. Плауновидные (Lycophyta)	4	236	4 (#1,#2,#3#4)
Класс Папоротникообразные (Polypodiopsida)	9	253	1 (#1)
Класс Гнетовидные (Gnetopsida)	2	27039	2 (#1,#2)
Класс Гинкговые (Ginkgoopsida)	1	21357	3 (#1,#2,#3)
Класс Саговниковые (Cycadopsida)	1	2922	2 (#2,#3)
Сем. Кипарисовые (Cupressaceae) и Сем. Тисовые (Taxaceae) - клада Conifers II	4	343330	2 (#1,#2)
Сем. Хвойные (Pinaceae) - клада Conifers I	9	372339	2 (#2,#3)
Класс Покрытосеменные (Magnoliopsida), древние группы - Порядки Амбореллоцветные (Amborellales), Кувшинкоцветные (Nymphaeales) и Магнолииды (Magnoliidae)	9	31343	1 (#1)

Элементы, содержащие добавочный домен aRNH, не были обнаружены среди всех исследованных геномов зелёных и стрептофитовых водорослей, а также маршанциевых мхов и наименее развитых таксонах настоящих мхов. Это подтверждает прежние предположения, что процесс приобретения добавочного домена произошёл у растений с уже сформировавшейся дифференцировкой тканей [9]. 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 [9].

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 отражены на рисунке 12, а в таблице 2 отражено количество структур, выявленных в каждом таксоне.

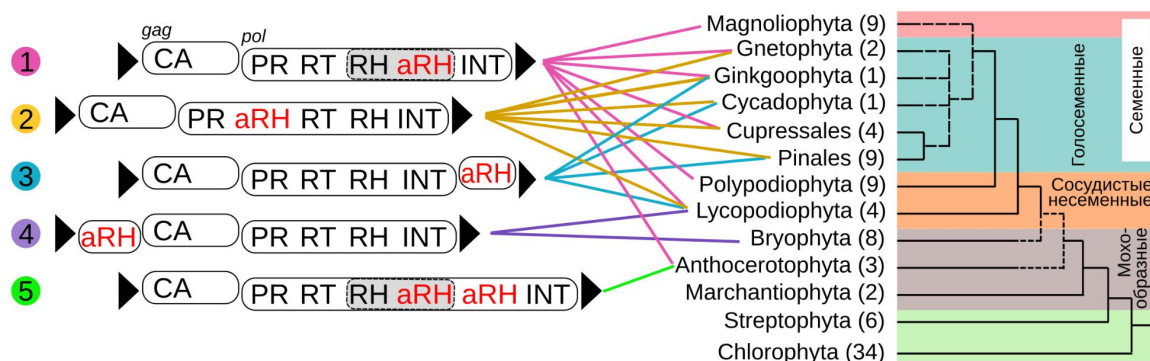


Рисунок 12. Схематичное изображение пяти доменных структур элементов *Tat*, различающихся по положению домена aRNH, и таксонов растений, в которых они были выявлены. Нумерация структур на изображении (1-5) соответствует дальнейшей нумерации в тексте (#1-#5). Схематическое представление филогении основных таксонов растений, представленное справа, базируется на серии публикаций [143–146]. Цветные линии соединяют структуру с таксонами, в которых она была выявлена. 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 - интеграз, RH или RNH - рибонуклеаза H, aRNH или aRNH - дополнительный домен рибонуклеазы H.

Структура #1, или “ретровирусная”, имеет последовательность доменов “GAG.PR.RT.RNH.aRNH.INT” (Рисунок 12). Единственная на данный момент структура с двумя доменами RNH (нативный и aRNH), выявленная в других группах LTR-ретротранспозонов за пределами кластера *Tat* (ретровирусы позвоночных и LTR-ретротранспозоны оомицетов *Chronos* и *Archon*). В данной работе элементы *Tat* со структурой #1 распространены среди наибольшего количества таксонов: антоцеротовые мхи, плауны, папоротники, древние голосеменные (гинкговые, гнетовые, тисовые и кипарисовые) и покрытосеменные.

Структура #2, “GAG.PR.aRNH.RT.RNH.INT”, впервые выявлена в прежней работе в геномах хвойных голосеменных и плауна *S. moellendorffii* [9] (Рисунок 12). В данной работе была показана высокая численность элементов данной структуры в геномах всех исследованных групп голосеменных, однако практически все они располагались в пределах двух кластеров, выявленных в

прежних работах. Исключение составила новая видоспецифическая линия с данной структурой, состоящая из элементов тисового генома *Taxus wallichiana*, которая произошла от линии со структурой #1 этого же вида (описанный случай рекомбинации можно проследить на рисунках в приложениях 2 и 4).

Структура #3, “GAG.PR.RT.RNH.INT.aRNH”, ранее была обнаружена только у *Tat* из геномов хвойных голосеменных [9] (Рисунок 12). Было показано, что в данной структуре домен aRNH расположен отдельной рамкой считывания в 3'-области элемента (прежде считалось, что домен остаётся частью гена *pol* - отражено на рисунке 7). В данном исследовании была расширена выборка геномов хвойных, у которых были найдены элементы *Tat* с данной структурой. Интересно, что данная структура была выявлена в невысокой численности в геноме древних голосеменных *Ginkgo biloba* и *Cycas panzhihuaensis*, а также плауна *Selaginella kraussiana*, элементы которых попали в два новых кластера *Tat* (Приложение 2).

Структура #4, “aRNH.GAG.PR.RT.RNH.INT”, выявлена в данной работе впервые (Рисунок 12). Домен aRNH расположен подобно структуре #3 вне генов *pol* и *gag*, но в 5'-области элемента, перед геном *gag*. Обнаружена среди элементов *Tat* из геномов плаунов и настоящих мхов.

Структура #5, также обнаружена впервые, представляет особый интерес - её элементы содержат три домена RNH: “GAG.PR.RT.RNH.aRNH-1.aRNH-2.INT” (Рисунок 12). Дальнейший анализ (описан ниже, в пункте “Сравнение степени деградации нативного домена RNH после приобретения aRNH”) показал, что домен aRNH-2 деградировал и в некоторых случаях распознавался только как принадлежащий к семейству RNH. Выявлена только у элементов *Tat* антоцеротовых мхов.

Таким образом, обнаруженные элементы подтвердили существование трёх выявленных прежде структур. Ранее, по содержанию элементов с этими структурами выделяли четыре кластера *Tat* из прежних работ [9,118,120]. Эти “старые” кластеры назывались *Tat* I, *Tat* II, *Tat* III, *Tat* IV-VI (составной кластер элементов с “ретровирусной” структурой с различными добавочными доменами в 5' и 3' UTR-областях; домены отражены на рисунке 7). Выявленные в ходе данной работы элементы расширяют многообразие *Tat*, что приводит к потребности отказаться от прежних обозначений и ввести новые (представлены

далее, в параграфе “Кластеры элементов *Tat* на основе филогении по домену RT”.

Кластеры элементов *Tat* на основе филогении по домену RT

Первоначально инструментом для построения быстрых деревьев FastTree были произведены филогенетические реконструкции по аминокислотным последовательностям трёх основных доменов LTR-ретротранспозонов: RT, INT и нативному домену RNH. На каждой из реконструкций наблюдалось единообразие по сформированным кластерам элементов. Имелись незначительные пертурбации с положением групп относительно друг друга, не искажающие общую картину. Конкатенированное выравнивание трёх последовательностей было передано на повторную филогенетическую реконструкцию инструментом IQTree. Режим автоматической подборки эволюционной модели и критериев реконструкции выбрал режим “LG+F+G4”. Итоговое филогенетическое дерево конкатенированных последовательностей выглядело полностью соответствующим дереву по домену RT. Поэтому в дальнейшем, для простоты описания, будет упоминаться как дерево по домену RT. Схематическое изображение дерева представлено на рисунке 13. Совокупность “старых” и впервые выявленных кластеров внутри *Tat* обозначены латинскими буквами *A*, *B* - *L*, выделенными курсивом.

Весь кластер *Tat* можно разбить на две ветви. В первой, одним из наиболее базальных кластеров расположен (отображено на рисунке 13) кластер *Tat Z* (“Zero”). Элементы этого кластера, по всей видимости, состоят из ранних элементов *Tat*, существовавших до приобретения домена aRNH и деградации нативного RNH. Этот кластер, названный в прежних работах кластером *Tat*-подобных элементов мхов (“moss-*Tat*-like” на рисунке 7) [9,118] выявляется только поиском через домен RT. В данный кластер попадают элементы всех таксонов растений от трёх групп мхов до голосеменных включительно. Кластер *Tat Z* не представляет большого интереса с точки зрения данной работы, кроме описанного ниже к пункту “Сравнение степени деградации нативного домена RNH после приобретения aRNH”.

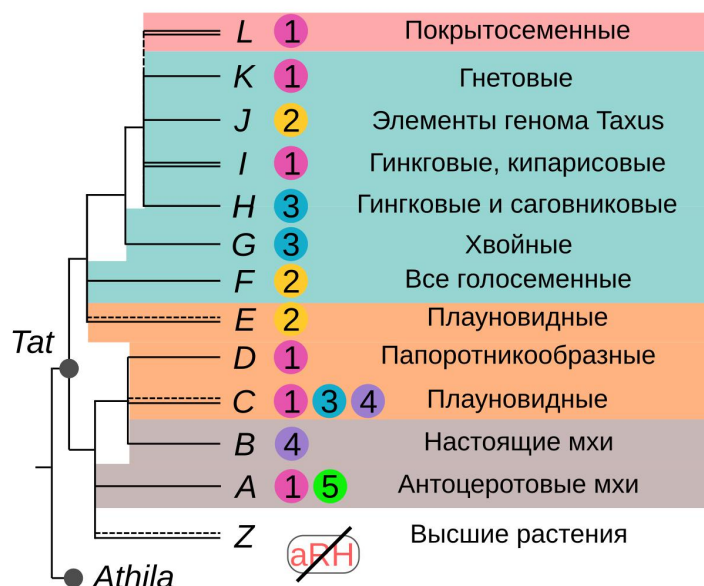


Рисунок 13. Схематическое филогенетическое древо взаимоотношений *Tat* LTR-ретротранспозонов, построенное методом максимального правдоподобия на основе аминокислотных конкатенированных последовательностей доменов RT, RNH, INT, и выполненное в виде кладограммы с условными длинами ветвей. Консенсусные 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 *Athila*) ответвляются кластеры *Tat*, 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.

Остальная часть первой ветви *Tat* (кластеры *Tat A-D, Z*) целиком состоит из кластеров, выявленных впервые в рамках данной работы (Рисунок 13).

Элементы именно этой ветви представлены новыми исследованными таксонами нецветковых растений, и именно в них обнаружены две новые структуры элементов с aRNH. Ветвь может быть рассмотрена как состоящая из двух ключевых узлов: (1) базального, откуда отходят вышеупомянутый кластер *Tat Z* и кластер *Tat A*, представленный геномами антоцеротовых мхов, а также (2) следующего за ним узла, где без достоверной поддержки ветвления расходятся друг от друга три остальных кластера (*Tat B, C, D*).

Кластер *Tat A* представлен элементами двух структур (#1 и #5) из геномов антоцеротовых мхов (Рисунок 13). Примечательно, что, будучи расположенными “вперемежку” внутри кластера *Tat A*, обе структуры могут восприниматься как производная одна от другой.

Кластер *Tat B* представлен элементами геномов настоящих мхов со структурой #4 (Рисунок 13). Только в половине исследованных геномов мхов были обнаружены элементы с aRNH, что может быть объяснено как неполнотой сборки некоторых геномов, так и некими эволюционными процессами, происходившими в отделе настоящих мхов (Таблица 2, приложение 1). Такими процессами могут быть различные преобразования геномов или изоляция, не позволившая отдельным линиям приобрести как сам домен aRNH, так и копию уже образовавшегося элемента путём горизонтального переноса. Для изысканий в данном вопросе наши данные не предоставляют достаточных сведений.

Кластер *Tat C* полифилетичен и представлен разнообразием элементов различных структур из геномов плаунов (Рисунок 13). Стоит отметить, что это один из двух полифилетических кластеров, сформированных элементами геномов плаунов, выявленный впервые в рамках данной работы. Всего в данном кластере имеется 3 структуры LTR-ретротранспозонов *Tat* с доменом aRNH (#1, #3 и #4). Все линии данного кластера “перемешаны” на древе, что не позволяет судить о взаимоотношении между элементами с этими структурами. Из четырёх исследованных геномов плаунов в составе данного кластера имеются элементы только двух, относящихся к отделу Селагинелловых (*Selaginellales*), одной из трёх основных групп класса Плауновидные (Таковыми являются: отдел Полушниковые или *Isoetales*, отдел Плауновые или *Lycopodiales*, отдел Селагинелловые *Selaginellales*. В таблице 1 и приложении 1 классификация представлена более высокими рангами).

Кластер *Tat D* сформирован элементами папоротников со структурой #1 (Рисунок 13). Элементы не были обнаружены лишь в некоторых “черновых” геномных сборках. Рассуждать о причинах отсутствия элементов *Tat* в данных геномах с имеющимися данными не представляется возможным.

Вторая ветвь *Tat* состоит преимущественно из выявленных ранее кластеров, называвшихся *Tat* I-VI [9,118] (Рисунки 7, 13). Однако, в силу обнаружения среди них ещё нескольких внутренних кластеров (между *Tat* III и IV), прежняя нумерация перестала быть удобной. В новых обозначениях мы предлагаем делить ветвь на 8 кластеров *Tat E, F, G, H, I* (не путать с *Tat* I - прежняя нумерация по римскому образцу не выделяется курсивом), *J, K* и *L*.

Кластер *Tat E* (прежний *Tat* I) со структурой #2 ранее был выявлен только в геноме плауна *Selaginella moellendorffii* в количестве трёх элементов [9] (Рисунок 13). Теперь около 50 копий обнаружены в том же геноме и 1 копия в геноме другого плауна - *Selaginella kraussiana*. Вдобавок, у *Isoetes engelmannii*, представителя другой большой линии плаунов, класса Полушниковых (Isoetopsida), также выявлена данная структура, однако, его положение неоднократно падало либо в виде базальной ветви этого кластера, либо базальной ветви кластера *Tat F* (прежний *Tat* II с рисунка 7). Для простоты представления, в данной работе эти элементы будут рассматриваться как представители одного полифилетического кластера.

Кластер *Tat F* (прежний *Tat* II) с аналогичной структурой (#2), свойственный голосеменным растениям, также был выявлен с невысокой численностью в прежних работах [9] (Рисунок 13). В проделанной работе были выявлены элементы, попавшие в данный кластер, количеством в десятки тысяч элементов на геном. Кластер по-прежнему представлен только элементами геномов голосеменных, как древних и практически вымерших таксонов (гнетовых, *Gnetum montanum*, гинкговых, *Ginkgo biloba*, и саговниковых, *Sycas panzhihuaensis*), так и представителей семейств хвойные, тисовые и кипарисовые (Таблица 2, Приложение 1).

Кластер *Tat G* (прежний *Tat* III) со структурой #3 ранее был выявлен в небольшом количестве у хвойных и вызывал вопрос о положении домена aRNH (Рисунок 13). Так, в работе [9] домен изображён, как С-концевой в полипептиде гена *pol* (Рисунок 7). Хотя полученные в ходе данной работы элементы, попавшие в данный кластер, обладали той же последовательностью доменов,

поиск ОРС (детальнее разобран ниже в пункте “Поиск добавочных ОРС с кодирующими последовательностями”) показал, что домен aRNH расположен в отдельной ОРС в 3' области, примерно в 500-1000 нуклеотидах от окончания ОРС гена *pol*. Новые данные также расширили представленность кластера *Tat G* новыми геномами, по-прежнему принадлежащими только к хвойным голосеменным (Таблица 2, Приложение 1).

Кластеры *Tat H*, *I*, *J* и *K* выявлены впервые между *Tat G* (прежний *Tat III*) и *Tat L* (прежний *Tat IV-VI*) (Рисунки 7, 13).

Кластер *Tat H* представлен элементами двух родственных друг другу древних групп голосеменных - гингковых и саговниковых (также известных как цикасовые). Элементы кластера представлены структурой #3, как и элементы ближайшего к ним кластера *Tat G* (Рисунки 7, 13).

Кластер *Tat I* представлен элементами со структурой #1 из геномов *G. biloba* и всех четырёх исследованных видов кипарисовых (Рисунки 7, 13). Элементы гинкго лежат в основании кластера, длины ветвей не позволяют судить о сильной удалённости данных элементов от кипарисовых.

Кластер *Tat J* технически не является оформленным кластером - он представлен элементами тисового генома - *Taxus wallichiana* - со структурой #2. При этом, для данного генома по-прежнему обнаруживаются элементы с аналогичной структурой в кластере *Tat F*, содержащем элементы всех групп голосеменных с #2. С другой стороны, кластер *Tat J* периодически определяется на древе RT как внутренний в составе *Tat I*, внутри которого находятся элементы этого же генома кипарисового со структурой #1 (Рисунки 7, 13).

Кластер *Tat K* представлен элементами гнетовых - *Gnetum montanum* и *Welwitschia mirabilis* со структурой #1 (Рисунки 7, 13). Представляет интерес таксоноспецифичность данного кластера по отношению ко всем кластерам с аналогичной структурой, в особенности к кластеру *Tat K*, в котором смешались элементы со структурой #1 от разных групп голосеменных.

Кластер *Tat L* (прежние *Tat IV-VI*) со структурой #1, согласно прежним исследованиям, представлен геномами покрытосеменных [9] (Рисунки 7, 13). Изначально в рамках данной работы не планировалось рассматривать элементы покрытосеменных из данной группы. Недавно стали доступны геномы древних групп покрытосеменных - порядки Амбореллоцветные (Amborellales), Кувшинкоцветные (Nymphaeales) и Магнолииды (Magnoliidae), отражённые в

таблице 1. Предполагалось, что у древних таксонов покрытосеменных могли сохраниться структуры, отличные от структуры #1. Однако, хотя геномы этих растений сформировали новые клады внутри *Tat I*, все элементы этой группы по-прежнему были представлены структурой #1. Для простоты представления в рамках данной работы, кластер по-прежнему будет отражаться единым, хотя на деле представлен несколькими таксон-специфическими группами.

Помимо систематически повторяющихся структур, сформировавших линии с высокой численностью, представленных на рисунке 13, были также обнаружены малочисленные синглетные последовательности элементов со структурами, которые не выявлялись систематически в соответствующих таксонах. Их анализ, объясняющий невовлечённость в данное филогенетическом древе, производится в блоке дискуссии, главе “Анализ спорных элементов-синглетов”.

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 14. Полное древо приведено в приложении 2. В глобальном плане, оно повторяет филогенетическое древо (приведено в приложении 3) доменов RNH из прошлого исследования, т.к. содержит те же три основные ветви.

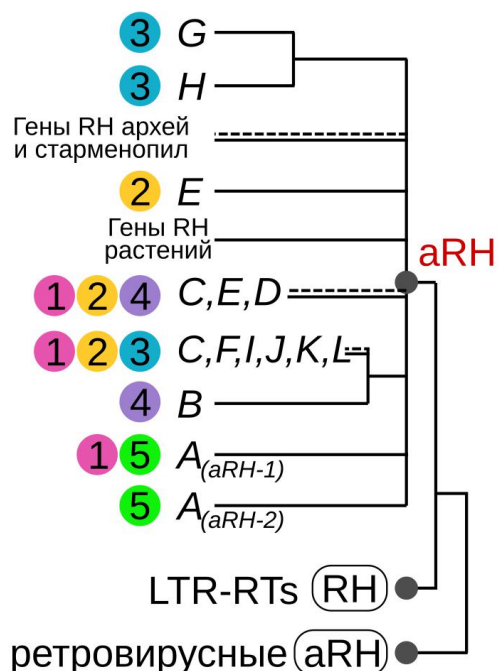


Рисунок 14. Схематическая визуализация филогенетического дерева доменов RNH с акцентом на кластера доменов aRNH, выявленных преимущественно в элементах *Tat* (узлы, не имевшие достоверных значений, объединены до ближайших базально расположенных к ним достоверных узлов; длины ветвей условны). Среди трёх основных эволюционных ветвей домена рибонуклеазы H (**ретровирусные aRNH** - 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, **aRNH** - гены RNH растений, страменопил и архей). Оригинальное дерево укоренено кластером доменов **ретровирусных aRNH**. В кластере **aRNH** 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 13. Кластеры выделены цветами по соответствию структурам на рисунке 12: розовый - #1, желтый - #2, синий - #3, фиолетовый - #4 и зелёный - #5. Одновременно прямой и пунктирной линией обозначены полифилетические кластеры.

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 несколько кластеров (Рисунок 14). Взаимоотношение этих кластеров между собой остаётся спорным, поскольку многие узлы не прошли выбранного критерия достоверности (поддержка aLRT и ufBoot - 80 и 95, соответственно) и можно утверждать, что все они выходят из одного узла. 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 мутирования самих ретротранспозонов, недостаточной, однако, для того, чтобы элементы могли быть достоверно разделены на разные группы.

Филогения по домену aRNH (Рисунок 14) во многом соответствует филогении по домену RT (Рисунок 13). Так, для 4 из 5 кластеров несеманных растений с древа RT (*Tat A*, *C*, *D* и *E*, рисунок 13) сформированы практически соответствующие им кластеры. Для двух полифилетических кластеров плаунов с древа RT (Рисунок 13) был получен один большой полифилетический кластер aRNH, содержащий элементы из геномов всех исследованных представителей рода *Selaginella*, все элементы папоротников, выделенных внутри него в отдельную ветвь, а также две линии aRNH плаунов расположились отдельно: одна независимо (aRNH из генома *I. engelmannii*, *Tat E*), другая линия (aRNH из элемента со структурой #3 генома *S. kraussiana*) попала в большой кластер семенных растений (Рисунок 14). Стоит отметить, что элементы плаунов располагаются на реконструкциях древа RNH крайне нестабильно: иногда aRNH из генома *I. engelmannii* и aRNH из генома *S. kraussiana* со структурой #3 группируются с другими aRNH плаунов, однако, в таких случаях распадается большой кластер aRNH плаунов, оставляя aRNH папоротников независимым монофилетичным кластером. Подобная непостоянность может говорить в пользу незначительного эволюционного расстояния между кластерами.

Для антоцеротовых мхов выявлено два кластера (Рисунок 14) - основной с проксимальными aRNH (единственный aRNH для структуры #1 и aRNH-1 из

элементов со структурой #5), и второй с “лишним” aRNH (aRNH-2) из структуры #5 с тремя доменами RNH (нативный RNH, aRNH-1, aRNH-2).

Элементы кластеров семенных растений с древа RT (*Tat F, G, H, I, J, K* и *L*, рисунок 13) сформировали два соответствующих им кластера на древе aRNH (Рисунок 14). В один из этих кластеров попали домены aRNH элементов со структурой #3 из кластеров древа RT *Tat G* и *H*, в другой - все остальные домены aRNH элементов перечисленных кластеров. Туда также попали домен aRNH из элемента плауна *S.kraussiana* со структурой #3 (расположился внутри ветви) и элементов настоящих мхов (*Tat B*), достоверно расположившиеся базальной ветвью ко всему кластеру.

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 15.

Structure	RH				aRH			
	D	E	D	D	D	E	D	D
1	X	X	X	X	✓	✓	✓	✓
2	X	✓ X	X	X	✓	✓	✓	✓
3	X	X	X	X	✓	✓	✓	✓
4	X	✓ X	X	X	✓	✓	✓	✓
5	X	X	X	X	✓	✓ X	✓	✓
по aRH	✓	✓	✓	✓	X	X	X	X

Рисунок 15. Зависимость деградации нативного домена RNH от наличия домена aRNH в различных положениях для выявленных структур по положению домена aRNH. Столбец “Структура” отражает пять структур (#1-#5) с aRNH или исходную структуру без добавочного домена aRNH. Справа от столбца “Структура” в табличном виде изображены 8 столбцов по 4 на каждый домен (нативный RNH и aRNH, на рисунке домены обозначены “RH” и “aRH”, соответственно), отражающих наличие в последовательности элемента на соответствующих позициях аминокислот каталитически активного центра тетрады DEDD.

Основным наблюдением в этом свете 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 (включая наиболее приближённые к *Tat* с aRNH - *Athila* и *Tat Z*) 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.

Анализ спорных элементов-синглетов

Поиском командой 'grep' по файлу в командной строке (bash) выявлялись единичные элементы со структурами, не выявленные в каждом отдельном геноме систематическим поиском. Так, количество потенциальных структур, не вошедших в анализ представлено в таблице-приложении 5. Среди них был произведён частичный отсев по очевидно химерному происхождению структур (те случаи, где в аннотации структуры участвуют лишние домены, и алгоритм не смог разбить эти последовательности на два элемента). Затем, после беглого осмотра последовательности элемента в онлайн-инструменте CD-search, оставшиеся элементы были добавлены к выравниваниям двух деревьев (конкатенированное древо RT+RNH+INT и древо RNH-aRNH) инструментом mafft. Полученные деревья содержали 61 элемент из 69 предсказанных элементов (8 забракованы программным обеспечением для построения филогенетических деревьев). Сравнение данных деревьев приводится в приложении 6.

Среди этих элементов были элементы плауна *Isoetes engelmannii* со структурой #1, плауна *Selaginella tamariscina* со структурой #3, элементы гнетового *Welwitschia mirabilis* со структурой #2, элементы хвойных со структурами #1, #4 и кипарисовых с #3, #4.

В случае элементов плауна *Isoetes engelmannii* со структурой #1 на каждом из деревьев домены соответствовали кластеру за пределами *Tat*, свойственному паразитам растений - оомицетам (кластер *Chronos*). Поскольку геномная сборка не доведена до хромосом, а количество контигов представлено десятками тысяч последовательностей, нам видится наиболее вероятным объяснение о

контаминации образца, нежели вероятность горизонтального переноса. В приложении 6 данные элементы группируются в один кластер на каждом из деревьев, подписаны на древе aRNH '*Chronos*'.

В случае элементов *Welwitschia mirabilis* со структурой #2, все пять элементов попали в кластер *Tat F* на древе RT, а их aRNH - в кластер aRNH семенных растений со структурами #1 и #2. Таким образом, не выявленная в больших количествах, структура всё же не была утрачена у отдельного генома гнетовых, как приходилось бы предполагать без данного анализа. В приложении 6 данные элементы группируются в один кластер на каждом из деревьев, подписаны на древе RT '*Wmir #2*'.

В случае остальных структур наблюдался один из двух сценариев: (1) элементы не соответствовали элементам *Tat* по одному или двум деревьям; (2) элементы соответствовали каждому из филогенетических деревьев, но не могут однозначно считаться не случайными структурами. Так, большинство из оставшихся элементов либо не соответствовали кластеру *Tat* по домену RT (и двумя другими, конкатенированными с ним), либо домен aRNH этих элементов лежал в кластерах, соответствующих кластерам aRNH растений-хозяев или в кластерах одиночных aRNH. Происхождение последних дискуссионно - они не тяготеют к кластерам aRNH *Tat*, но и лежат отдельно от других доменов растений. Мы предполагаем, что эти домены сохранились после деградации древних элементов *Tat*, обнаружить которые уже не представляется возможным.

К единичным случаям, в которых по каждому из деревьев оставшиеся синглетные элементы, стоит отнестись осторожно: они лежат внутри кластера *Tat*, их aRNH соответствует одному из кластеров aRNH *Tat*, однако, расположились они не в кластерах с аналогичными структурами. Из этого следуют три вероятные интерпретации - (1) химеризм (но, в отличие от отброшенных по несоответствию с одним из деревьев, эти химерные элементы сформированы двумя элементами *Tat*); последние сохранившиеся элементы в ходе элиминации; или же результат рекомбинации внутри кластера. В последнем случае элемент, недавно прошедший через рекомбинацию должен подавать некоторые признаки целостности и автономности (практически 100% идентичность последовательностей LTR, целые домены), однако, этого не наблюдается. В противном случае, эти элементы стоит также рассматривать как "химерные". Рассматривая эти элементы, как сохранившиеся остатки от некогда

существовавших у предков элементов с данным структурами, следует обратить внимание, что в данном случае они не формируют единого кластера, а разбросаны по другим кластерам, структурно им не соответствующим. В данном контексте гипотеза об элиминации уступает химерному происхождению большинства описанных элементов.

Поиск добавочных ОРС с кодирующими последовательностями

Поиск добавочных ОРС онлайн-инструментами ORFfinder [141] и CD-search [127] (на основе базы данных CDD [147]) в кластерах таксонов несеманных и древних голосеменных растений ничего не дал.

Поиск дОРС выявил несколько систематически встречающихся доменов в геномах хвойных: SMC (мотив сегрегации хромосом, функция неизвестна), CCDC (мотив coil-coiled, обнаруживался в ранних работах [9]), PNA03247 (фрагмент в ~250 аминокислот белка длиной в >3000 аминокислот; выявлен не систематически - по несколько случаев на кластер), COG5283 (фрагмент белка, согласно аннотации свойственного мобилому, встретился единожды) и множество других фрагментированных доменов (PAT1, Med15, PRK, PABP) в 5' области и SMC в 3' области. Причина фрагментированности этих доменов не ясна, но картина всегда одна: обнаруживается захваченной лишь часть от длины домена, известной по базе данных CDD. Захват этих фрагментов мог происходить как случайно и без какой-либо пользы для элемента, так и с целью "кусочного" захвата больших доменов ради частичного использования их свойств. Также эти структуры могут оказаться не изученными прежде доменами, которые оказались в какой-то степени близкородственными или конвергентными известным.

дОРС не были обнаружены в древних геномах голосеменных из групп гинкговых и гнетовых, но у кипарисовых более чем в половине исследованных элементов был расположен plant mobile domain (PMD, функция неизвестна) в 3' области, а также встречаются немногочисленные фрагментированные домены PNA03247, PAT1, MMR1 в 5' области. Фрагмент домена PNA03247 также был обнаружен в геноме саговниковых.

Вопрос о внутренней дивергенции кластеров *Tat* по приобретению этих добавочных фрагментированных доменов требует большего внимания, поскольку представленная выборка не может считаться достаточной.

Вышеописанные данные представляют всего по 25 наиболее сохранившихся элементов каждой линии на кластер, в то время как численность некоторых линий достигает десятков тысяч элементов. Также представляет интерес возможная роль таких добавочных “фрагментированных” доменов.

Поиск добавочных ОРС также подтвердил, что в элементах со структурами #3 и #4 (рисунок 12) домен aRNH располагался в отдельных от генов *gag* и *pol* рамках считывания (3' - и 5' - области соответственно).

Глава 4. Обсуждение результатов

Захват добавочного домена aRNH LTR-ретротранспозонами наземных зелёных растений сыграл весомую роль в распространении элементов *Tat*. В его результате появилось высокое разнообразие структур, по всей видимости, отличающихся друг от друга разной степенью адаптивности. Однако, имеющихся данных по-прежнему не достаточно, чтобы ответить на некоторые из вопросов, касаемых данного эволюционного события.

Так, остаётся не ясно, имеет ли эволюция элементов *Tat* с добавочным доменом aRNH однонаправленной, как это было в случае ретровиурсов позвоночных и LTR-ретротранспозонов оомицетов, или же имеется иное объяснение отсутствию иных структур в данных группах. Неизвестна причина, по которой домен aRNH был приобретён. Было это случайностью или эссенциальной необходимостью для сохранения автономности элементов *Tat*? Был ли процесс захвата единократным или происходил неопределённое множество раз? Как сменяли друг друга эти структуры в ходе эволюционного процесса? Является ли многообразие выявленных в *Tat* структур результатом постоянных рекомбинаций или же их частота была незначительной? Имел ли место горизонтальный перенос мобильных элементов данной группы среди геномов зелёных растений или же их эволюция шла вертикально?

Чтобы понять это, необходимо рассмотреть случаи несоответствия данных филогенетических деревьев RT и RNH, а также провести параллели с эволюцией геномов-хозяев. Рассмотрение процесса деградации домена RNH способно дать ответы о причинах захвата добавочного домена RNH. Рассмотрение обнаруженных структур и потенциально возможных может дать намёк на понимание, в каком направлении движется эволюционный процесс данной группы LTR-ретротранспозонов.

Несоответствия деревьев RT и RNH

Основными несоответствиями между филогенетическими деревьями по доменам RT и aRNH являются:

- Близкое расположение aRNH элементов настоящих мхов (*Tat B*) к aRNH большинства кластеров семенных растений древа RT (*Tat F, I-L*).

- Удалённость доменов aRNH элементов семенных со структурой #3 (*Tat G, H*) от остальных aRNH семенных (*Tat F, I-L*).
- Оба “кластера” aRNH плаунов содержат домены каждого из двух кластеров древа RT (*Tat C* и *E*).

Эти три несоответствия можно объяснить быстрой дивергенцией весьма коротких последовательностей, что не позволяет детальнее разобрать и объяснить природу наблюдаемых взаимоотношений доменов aRNH. При рассмотрении полученного количества кластеров с прежней позиции неоднократных захватов aRNH, пришлось бы утверждать о минимум 6 случаях захвата.

Однако, есть примеры того, как (1) элементы удалённых кластеров *Tat* (*F* и *I-L*, Рисунок 13) попадают в общий кластер aRNH (ветвь *C-F-I-J-K-L*, Рисунок 14), или (2) элементы двух удалённых кластеров *Tat* (*C* и *E*, Рисунок 13) на древе aRNH распределяются преимущественно в составе одного кластера aRNH (ветвь *C-D-E*, Рисунок 14), в то время как две оставшиеся линии разбросаны хаотично. Эти примеры говорят о, вероятно, незначительном расстоянии между кластерами aRNH древа RNH, что, в свою очередь, сокращает количество возможных случаев захвата aRNH и ставит под вопрос филогенетическое расстояние между всеми 6 кластерами aRNH. Более того, это наводит на мысль о том, что все связанные данными примерами кластеры могли дивергировать от одного источника.

Таким образом, эти несоответствия вместе с недостоверными коэффициентами все же, более вероятно, говорят в пользу единичного захвата элементами *Tat* домена aRNH, в отличие от предполагавшейся ранее серии независимых захватов элементами *Tat* данного домена.

Ещё одним несоответствием двух филогенетических деревьев (Рисунки 13, 14) является наличие двух кластеров aRNH у антоцеротовых мхов. Однако, данное несоответствие объясняется тем, что только у антоцеротовых мхов была выявлена структура #5, где содержалось 3 домена RNH - нативный RNH, aRNH-1 и aRNH-2. aRNH-1 попал в один кластер с элементами aRNH антоцеротовых мхов со структурой #1. Домены aRNH-2 сформировали второй кластер с достоверной поддержкой. Взаиморасположение этих кластеров говорит о том, что домены aRNH-1 и aRNH-2 достаточно близкородственны, и в ходе

возникновения структуры #5 наиболее вероятно имели место два события: дубликация домена aRNH и последующая деградация одного из них.

Сам факт деградации не только нативного домена RNH, но и одного из aRNH может косвенно говорить о конкуренции доменов с одинаковой функцией, приводящей к потере функциональности всех избыточных копий домена.

В прежних работах предполагалось, что такая дивергенция доменов aRNH кластера *Tat* объясняется независимым повторным захватом. В силу небольшой выборки формировались только три кластера на древе aRNH (Приложение 3), которые, в силу пертурбаций построения деревьев, располагались не последовательно рядом, а с промежуточными кластерами aRNH организмов, которые, как предполагалось, могли быть источниками захвата. В данной работе, благодаря увеличенной выборке, большинство сформированных доменами aRNH кластеров расположены последовательно. Коэффициенты поддержки между многими из 6 кластеров доменов aRNH имеют низкие значения, что говорит о быстрой дивергенции от потенциально одного общего предка. На древе (Рисунок 14) это отражено в схлопывании узлов в один общий. Таким образом, предположения о неоднократном приобретении домена aRNH [9,118] в ходе эволюции *Tat* уступают гипотезе о единичном приобретении с последующей дивергенцией.

Интересным случаем являются домены aRNH элементов генома *Taxus wallichiana*, которые повторно приобрели структуру #2 (*Tat J*, Рисунок 13) внутри ветви, состоявшей из кластеров элементов со структурой #1, а ближайшие кластеры, представленные не структурой #1, содержали структуру #3 (*Tat G, H*, Рисунок 13). На филогенетическом древе aRNH (Приложение 2) в рамках большого кластера aRNH семенных расположены три линии элементов этого генома. Одна, удалённая от других двух, соответствует элементам кластера *Tat F* с древа по домену RT (Рисунок 13). Остальные две (со структурами #1 и #2) располагаются рядом, как и на древе RT (приложении 4). Мы рассматриваем это происхождение повторно образовавшейся структуры #2 как результат нового раунда рекомбинации из близкородственной линии со структурой #1. Альтернативная гипотеза в том, что среди ранней линии *Tat I* сохранялись в малочисленном количестве элементы со структурой #2, эволюционировавшие как элементы со структурой #1 и дошедшие до наших дней лишь в одном геноме тисовых как *Tat J*, кажется менее вероятной.

Этот случай интересен как самим процессом конвергенции к уже существующей в голосеменных структуре #2 из другого кластера, так и процессами, которые могли вызвать гипотетическую рекомбинацию в данном конкретном виде.

Дегградация нативного домена RNН

Фундаментальный вопрос процесса дегградации одного домена RNН при наличии нескольких - “В какой последовательности происходили события дегградации нативного домена и приобретение дополнительного?” - остаётся нерешённым. Связь между двумя этими процессами видится из наших данных очевидной - не было выявлено элементов, которые могли бы являться промежуточными этапами в данном процессе. Такими гипотетическими интермедиатами могли быть элементы с целым активным центром нативного RNН при наличии домена aRNН или элементы без aRNН с дегградировавшим нативным доменом.

Нами рассматривались две гипотезы: (1) нативный домен дегградировал (приведя к потере элементом автономности), после чего элемент захватил домен aRNН, восстановив тем самым свою активность; (2) приобретение домена aRNН было случайным и привело к дальнейшей дегградации нативного домена RNН, поскольку два активных домена с одной функции могли мешать работе друг друга. Формально, небольшим аргументом в пользу второй гипотезы может послужить существование структуры #5, где один из доменов aRNН, как и нативный RNН, потерял аминокислоты каталитического центра, в результате чего в элементе остался лишь один активный домен рибонуклеазы Н. Но наблюдение, что в элементе может присутствовать одновременно только один работающий домен, не является достаточным, чтобы делать какие-либо заключения по исходному вопросу.

Вышеописанные наблюдения подтверждают предшествовавшие данные [9] (Рисунок 8) о дегградации нативного домена в *Tat*, показывая тенденцию, аналогичную той, что произошла в ретровирусах, а также в близкородственных им локи- и одинретровирусах (*lokiretrovirus* [148] и *odinretrovirus* [149]). “Старый” домен RNН постепенно теряет аминокислоты каталитически активного центра и, в случае “ретровирусного” (структура #1) положения aRNН

имеет перспективы превращения в домен-связку, как это произошло у ретровирусов позвоночных и LTR-ретротранспозонов оомицетов [9,115,118].

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 изначально происходил в данное положение 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.

Если делать допущение, что именно “ретровирусная” структура приводит к деградации нативного домена RNH, то она должна быть первой сформировавшейся после захвата aRNH структурой у *Tat*. Иначе бы сохранились произошедшие от неё структуры с целым нативным доменом. Это служит возможным объяснением многообразия таксонов растений с элементами *Tat*, представленных ей - структура #1 попала во многие из них практически сразу после захвата aRNH. Утрате аминокислот каталитически активного центра нативного домена RNH было бы объяснение аналогичное ситуации с ретровирусами, поскольку именно на двухдоменном “ретровирусном” положении была впервые показана деградация RNH [115] и сделано предположение, что именно при такой структуре происходит деградация нативного RNH [9]. Несмотря на сохранение вторичной и третичной структуры [150] во всех трёх группах - (1) LTR-RTs RNH (нативный RNH), (2) aRNH и (3) RNH ретровирусов, грибов и животных (fmRNH) - их активность сильно различается [8,115,151,152]. Известно, что нативный домен RNH не обладал одной из возможных дополнительных аминокислот активного центра (R или H в положении между последними двумя D тетрады DEDD) [115,118,120]. Это, а также накопление мутаций, сделало его менее каталитически активным, чем другие RNH типа I, такие как aRNH или fmRNH. Предполагается, что LTR-ретротранспозонам это было выгодно, поскольку при полной активности домен RNH был бы способен расщеплять РНК ретроэлемента прежде, чем начался бы процесс обратной транскрипции. Однако, подобного расщепления не происходит в ретровирусах, где было показано, что нативный домен, после деградации превратившийся в домен-связку, стал связующим звеном в едином пептиде из

доменов RT, домена-связки (прежний нативный RNН) и новый RNН (произошедший от fmRNН). Это объясняется тем, что домен-связка регулирует расстояние между новым RNН и RT, а также меняет конформацию (а с ней и активность) самого RNН ретровируса, чем достигается необходимый ретроэлементу специфический уровень активности рибонуклеазы Н [8,84,115]. Если описанная для ретровирусов картина аналогична для элементов *Tat*, то преимущество структуры #1 над остальными становится очевидным.

Но, если “ретровирусная” структура (#1) не была первой, а деградация RNН происходит вне зависимости от положения новоприобретённого домена, возникают вопросы: “Как сменяли друг друга эти структуры?” и “Почему остальные структуры сохранились и даже размножились?”.

Разнообразие структур в геномах несеменных таксонов растений, в этом случае, вероятно, является результатом “вспышки” рекомбинаций, в результате которой из неизвестной исходно захватившей aRNН структуры возникло разнообразие структурных вариантов. Те из них, что были наиболее благоприятными (в частности, структура #1) сразу же закрепились в некоторых таксонах (антоцеротовые мхи, плауны, папоротники), оставшись в некоторых из них практически единственными. Численное превосходство структуры #2, сохранившейся начиная с геномов плаунов, проявилось в явном виде на геномах голосеменных. Эта огромная копияность элементов на геном отражает некое преимущество структуры #2 над структурой #1, которое, вероятно, позволило этой группе элементов *Tat* размножиться в какой-то период эволюции растений. Разнообразие структур (#1, #2 и #3) на уровне таксонов *Tat F-L* может говорить о рекомбинации ради неизвестных ситуативных преимуществ структур и требует отдельного изучения.

Структуры: найденные, вероятные, производные и преобладающие

Среди всех выявленных в данной работе структур, четыре отличаются друг от друга по положению домена aRNН. При этом, среди всех гипотетически возможных положений, они представляются наиболее вероятными. Так, среди оставшихся позиций можно было ожидать домен aRNН внутри гена *gag*, в различных позициях между геном *gag* и доменом PR или между доменами RT и нативным RNН (Рисунок 16). Почему данные структуры не были обнаружены?

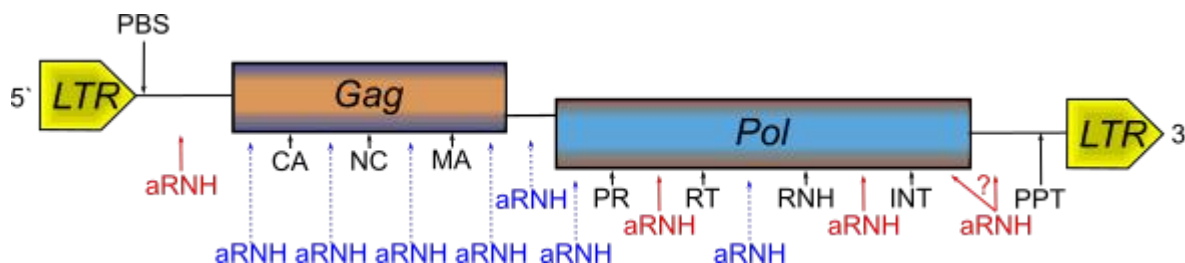


Рисунок 16. Обнаруженные и предположительно возможные структуры по положению домена aRNH *Tat* LTR-ретротранспозонов зелёных растений. На схеме представлены последовательности LTR, необходимые для транспозиции сайты и две OPC основных генов, свойственных LTR-ретротранспозонам. Стрелками черного цвета подписаны сайты (PBS, PPT) и домены. Стрелками красного цвета выделены положения домена aRNH, выявленные в этой или прежних работах [9,118,120]. Данные о включении в OPC гена *pol* домена aRNH, следующего за интегразным (INT) расходятся в прежней публикации [9] (в составе гена *pol*) и в текущей работе (отдельная OPC в 3' некодирующей области). Пунктирными стрелками синего цвета выделены положения домена aRNH, которые не были обнаружены, но, могут быть рассмотрены, как вероятные альтернативные позиции.

Известно, что белковый комплекс RT-RNH является единым немномерным продуктом гена *pol*. Поэтому встройка между этими доменами добавочного домена aRNH означала бы нарушение тонкой регуляции активности домена RNH за счёт конформационных особенностей его положения [110,117].

Домен PR отвечает за само-процессирование полипептида, получающегося из гена *pol*. Как уже разбиралось в обзоре литературы, этот процесс начинается с самовырезания с N-конца полипептида данным доменом, после чего освободившийся отдельный белок, состоящий из домена PR, обретая полную функциональность, завершает созревание продуктов *pol* [84,87]. Прежде уже было показано, что эффективность самовырезания PR и последующей за ней сборки зрелых вирусных частиц снижается при нарушении работы протеазы [153]. Поэтому, если бы домен aRNH оказался в составе *pol* до PR, у LTR-ретротранспозонов *Tat* с подобной структурой могли бы возникнуть сложности с самостоятельным процессингом белкового продукта гена. С другой стороны, если бы домен aRNH встроился как отдельная рамка считывания или как ещё один домен в составе *gag*, это, вероятно, нарушило бы регуляцию соотношения продуктов двух (или даже трёх) OPC. Подобные нарушения тратят лишние ресурсы клетки-хозяина, а потому эволюционно менее выгодны [153–156].

Встройка домена aRNH в составе гена *gag* в любом положении также видится вредоносным, а потому эволюционно невыгодным и маловероятным сценарием: участвуя в не связанных процессах жизнедеятельности

ретроэлемента домен aRNH способен нарушить регуляцию процессинга полипептидов с *gag*.

Всё вышеперечисленное видится нам вероятным объяснением, почему структуры с подобным положением aRNH не были выявлены, а также приводит к заключению, что обнаружение новых структур маловероятно.

Положение домена aRNH в структуре #3 (Рисунок 12) остаётся спорным. С одной стороны, проанализированные вручную элементы данной структуры содержали домен aRNH в виде добавочной OPC. С другой стороны, в силу высокой представленности элементов данной структуры (десятки тысяч копий), рассмотрение каждого элемента по отдельности вручную не представляется возможным. Поэтому существование структуры, где домен aRNH расположен в 3' области гена *pol* (сообщалось ранее в [9]), всё же допускается. В рамках данной работы этот вопрос не рассматривался, однако представляет определённый интерес.

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. Мы выдвигаем три гипотезы её вероятного происхождения.

Согласно первой, захват домена aRNH мог произойти из генома, где имелись две близкородственные друг другу копии aRNH (не в составе мобильного элемента). Они располагались близко (примерно в 500-1000 нуклеотидов) друг относительно друга и были захвачены в “ретровирусное” положение совместно сформировав элемент со структурой #5. После захвата один из доменов aRNH деградировал за невостребованностью, и, потенциально, мог породить структуру #1, выкинув из своего состава утративший функциональные свойства лишний домен aRNH (им оказался aRNH-2).

Согласно второй гипотезе, в элементах со структурой #1 начал деградировать первый приобретённый домен aRNH, и, для восстановления автономности, элементом был повторно приобретён ещё один домен aRNH (в положении между нативным RNH и старым aRNH, aRNH-2, поскольку деградировал именно он).

Согласно третьей гипотезе, трёхдоменная структура - результат дубликации домена aRNH внутри элемента, за которой последовала деграция одного из двух дубликатов.

Попытки выравнять дублицированные домены на нуклеотидном и аминокислотном уровнях инструментами BLASTn и PSI-BLAST не удалось в силу высокой дивергенции, однако, филогенетический анализ показал их высокое сродство, говорящее в пользу гипотез, где оба домена aRNH - результат дубликации на неизвестной стадии эволюции.

В свете имеющихся данных, нельзя однозначно сказать, какой структуре отдан эволюционный приоритет, как это можно судить в случае ретровирусов и LTR-ретротранспозонов оомицетов.

Чем же различаются эти структуры? Структуру #5 (Рисунок 12) можно рассматривать в данном контексте как структуру #1 с нефункциональным довеском aRNH-2. Структура #2 в рамках выборки геномов растений в данной работе показала себя самой многочисленной, копияность которой выше, чем у структур #1 и #3 в геномах голосеменных, являясь таким образом первой по численности и второй по распространенности среди выявленных структур (Приложение 7). В какой-то момент эволюции ранних семенных растений по неизвестной причине эта структура имела некое преимущество, позволившее ей многократно повысить свою копияность, сохранившуюся в геномах голосеменных до наших дней. Как и “ретровирусная” (#1), структура #2 содержит домен aRNH внутри гена *pol*, что как минимум гарантирует домену равную копияность с остальными доменами в *pol*. Возможно, она также, как RNH в ретровирусах не отрезается протеазой от домена RT и домена-связки, и её активность тоже как-то регулируется за счёт конформационных взаимодействий. Чтобы ответить на этот вопрос требуется исследование рибонуклеопротеинового комплекса RT-RNH с aRNH из разных положений физико-химическими методами. Без кристаллографического и белкового функционального анализа структуры данных комплексов не представляется возможным определить, какие преимущества даёт конкретное структурное положение домена aRNH. Возможно, такой анализ мог бы пролить свет на порядок возникновения различных структур *Tat* и причины закрепления разных групп в разных таксонах. Менее распространённые, чем структура #2, структуры #3 и #4 содержат домен aRNH в отдельных от гена *pol* ОРС, однако, многие их элементы проявляют

косвенные признаки недавней активности (идентичность LTR близкая к 100%, целые последовательности консервативных доменов - отражено в приложениях 5 и 7). Их численность по сравнению со структурами #1 и #2 относительно невысока. Из этого напрашивается предположение, что эволюционно данные структуры дают элементу меньшее преимущество из-за не достаточной синхронизации регуляции активности добавочных ОРС, как единственного видимого нам структурного отличия. Данное предположение также может быть подтверждено или опровергнуто за счёт кристаллографического и белкового функционального анализа.

Специфичность кластеров *Tat* таксонам растений

Полученные кластеры *Tat* специфичны по таксономическому составу (Рисунок 13). Сравнивая их с эволюцией растительных таксонов [143–146,157], стоит отметить преимущественно вертикальный характер эволюции элементов *Tat*.

Следующие исключения из таксоноспецифичной картины требуют дополнительного обсуждения:

- Удалённость кластеров плаунов на древе RT при относительной близости и перемешанности кластеров плаунов на древе aRNH.
- Разделение кластера *Tat E* на базовый *Tat E* (соответствует кластеру *Tat I* из прежних работ) и группу, тяготеющую к кластеру *Tat F*.
- Распределение структур #1 и #3 в геномах семенных.

Взаимоотношение кластеров плаунов с древа RT на наш взгляд стоит рассматривать со следующей позиции:

Ветвь *Tat A-D* (Рисунок 13), представлена наиболее древними таксонами, на уровне которых, предположительно, произошло приобретение aRNH. Ветвь *Tat E-L* (Рисунок 13) представлена кластерами семенных растений и одним кластером плаунов (*E*), расположенным базально. В силу наличия в геноме плаунов (*Selaginella kraussiana*) элементов из каждого кластера плаунов с древа RT (*Tat C* и *E*), можно предположить, что обе эти линии когда-то существовали одновременно в предковом плауне. В этом случае, возникновение обеих линий *Tat* могло произойти приблизительно одновременно. В этом случае, кластер *Tat E* может рассматриваться как часть первичного многообразия *Tat*, совместно с ветвью *A-D*, как *Tat A-E*. Тогда, кластеры *Tat C* и *E* соотносятся между собой по

структурному составу и эволюционному сосуществованию, как элементы кластера *Tat F* по отношению к кластерам *Tat G*, *H*, *I*, *K*, *L* голосеменных: как две группы линии (группа со структурой #2 против группы из структур #1 и #3) сосуществовали в одном или нескольких геномах, частично элиминируясь от группы к группе.

Кластер *Tat E* состоит из элементов 2 разных отделов плаунов - Полушниковых (Isoetales) и Селагинелловых (Selaginellales). Из них, именно элементы Полушниковых показали тенденцию тяготения к элементам голосеменных с аналогичной структурой. Поскольку считается, что голосеменные произошли от некоего древнего плауна, можно было бы предположить, что отдел Полушниковых был ближе к этому гипотетическому предку.

Вызывает интерес то, что среди всех семенных растений, исследованных в этой и прежних работах, только у современных голосеменных (хвойных) отсутствуют элементы со структурой #1. Напрашивается предположение, что они также содержали данную структуру, но по неизвестным причинам их эволюция отдала приоритет структурам #2 и #3 кластеров *Tat F* и *Tat G*, в то время как элементы со структурой #1 были утрачены. Альтернативной гипотезой, объясняющей представленные данные, может служить предположение, что эволюционно имела место рекомбинация, в ходе которой у остальных элементов домен aRNH переместился в “ретровирусное” положение, которая, однако, у хвойных не произошла.

От кластера древа RT (Рисунок 13), специфичного для покрытосеменных (*Tat L*), до кластера, специфичного для хвойных (*Tat G*) наблюдается интересная градация: элементы покрытосеменных *Tat L* (только структура #1), затем идёт кластер гнетовых голосеменных *Tat K* (#1), следом - кластер кипарисовых и гинкго (#1) - *Tat I*, саговниковые и гинкговые голосеменные (#3) - *Tat H*, а после - хвойные (#3) - *Tat G*. Расположенные между ними группы соответствуют основным группам растений, которые являются эволюционными ответвлениями между хвойными и покрытосеменными. При этом, рассматривая лишь по распределению структур #1 и #3, среди этих таксонов, и, соответственно, кластеров имеются геномы с только структурой #3 (*Cycas panzhihuaensis*), геном с обеими структурами (*Ginkgo biloba*) и геномы только со структурой #1, что

говорит в пользу гипотезы о наличии у неизвестного общего предка обеих структур, которые по-разному элиминировались от таксона к таксону.

Анализ филогении aRNH не даёт однозначного приоритета одной из данных гипотез, однако присутствие структуры #1 практически в каждом исследованном таксоне также позволяет предполагать, что элемент *Tat*, впервые захвативший домен aRNH, в краткие сроки сформировал все обнаруженные нами структуры, которые затем по-разному редуцировались в различных таксонах зелёных растений.

Всё многообразие структур семенных (*Tat F-L*), как эволюционно наиболее молодой группы растений, выявляется в одном из наиболее древних таксонов - гингковых. Являясь т.н. “живыми ископаемыми”, гингговые - одна из групп, наиболее близкая к эволюционному моменту дивергенции голо- и покрытосеменных, свидетельствуют в пользу наличия у общего предка семенных всех трёх структур, обнаруженных у современных семенных (структуры #1, #2 и #3). Все эти структуры уже были представлены в древних группах *Tat*, захвативших домен aRNH (*Tat A-E*), включая плаунов, от которых, согласно современным данным [143,145], произошли семенные растения. Это может говорить в пользу формирования однократного многообразия структур из некой первичной, которые затем наследовались, частично элиминируясь, от таксона к таксону. Однако, в силу высокой способности мобильных элементов к рекомбинации стоит также учитывать вероятность образования некоторых структур повторно. В силу представленности структуры #1 (“ретровирусной”) практически в каждом исследованном таксоне, нам кажется вероятным, что она образовалась первой, после чего дала начало остальным структурам.

Всё выше разобранное наводит на мысль о необходимости рассмотрения коэволюции элементов *Tat* и геномов растений. Она могла бы пролить свет на неразрешённые вопросы по эволюции растений (взаимоотношения мхов, радиация семенных, геномные делеции и дупликации) [157].

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 RNН 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 aRNН были обнаружены. Так, мы предполагаем, что это происходило в геноме некоего раннего предка всех групп мхов и плаунов. В силу невысокой выборки геномов Печёночных мхов нельзя однозначно утверждать, что эти события происходили после дивергенции от них. Также 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 aRNН с деградировавшим RNН, мы склоняемся в пользу гипотезы, где деградация нативного RNН следует за захватом aRNН.

Захват домена aRNН 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 aRNН 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 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.

Схематическое изображение расписанного выше эволюционного сценария представлено на рисунке 17.

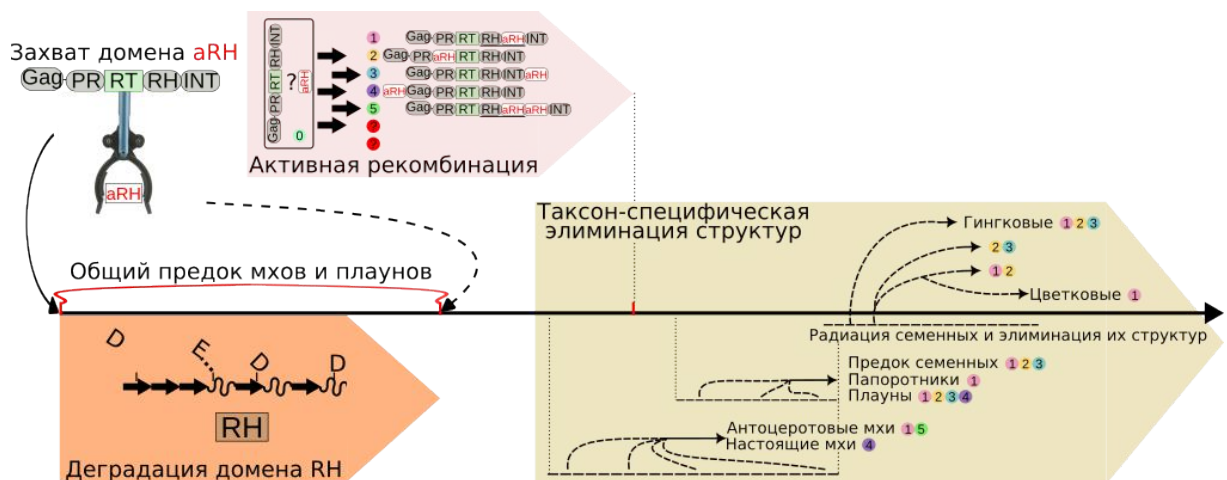


Рисунок 17. Предполагаемый сценарий эволюции *Tat* LTR-ретротранспозонов. Три ключевых события - захват добавочного домена aRNH (aRH), дегградация нативного домена RNH (RH) и активная рекомбинация - происходили в соответствующем порядке на уровне общего предка трёх линий мхов и плаунов. Данные события, по всей видимости, могли частично пересекаться во времени. В их результате у некого растительно организма сформировались как минимум 5 структур элементов *Tat*, различающихся положением домена aRNH. Далее, в силу неизвестных процессов, эти структуры элиминировались по-разному от таксона к таксону, пока не привели к обнаруженному многообразию.

Заключение по проделанной работе

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 видов с выявленными элементами и их численностью представлен в приложении 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 12).

LTR-ретротранспозоны *Tat* с доменом aRNH из геномов папоротников, плаунов, настоящих и антоцеротовых мхов сформировали на древе RT (рисунок 13) 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 (гинкгового *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 12, 13).

Филогенетический анализ домена aRNH (Рисунок 14, приложение 2) предполагает единственный случай приобретения домена aRNH у *Tat* LTR-ретротранспозонов с последующей дивергенцией.

Было проведено исследование зависимости деградации нативного домена рибонуклеазы Н от приобретения дополнительного домена aRNH (Рисунок 15). Связь этих двух событий была подтверждена, однако определить последовательность, в которой эти события происходили, остаётся невозможным. В силу особенностей выявленной структуры #5 (Рисунок 12), 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.

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 дублицированных 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.

Библиографические ссылки

1. Solyom S., Kazazian H.H. Mobile elements in the human genome: implications for disease // *Genome Med.* 2012. Vol. 4, № 2. P. 12.
2. Britten R.J. Transposable element insertions have strongly affected human evolution. // *Proc. Natl. Acad. Sci. U. S. A.* 2010. Vol. 107, № 46. P. 19945–19948.
3. Wicker T. et al. A unified classification system for eukaryotic transposable elements // *Nat. Rev. Genet.* 2007. Vol. 8, № 12. P. 973–982.
4. Eickbush T.H., Jamburuthugoda V.K. The diversity of retrotransposons and the properties of their reverse transcriptases // *Virus Res.* 2008. Vol. 134, № 1–2. P. 221–234.
5. Kapitonov V.V., Jurka J. Rolling-circle transposons in eukaryotes // *Proc. Natl. Acad. Sci.* 2001. Vol. 98, № 15. P. 8714–8719.
6. Vitte C., Bennetzen J.L. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. // *Proc. Natl. Acad. Sci. U. S. A. National Academy of Sciences*, 2006. Vol. 103, № 47. P. 17638–17643.
7. Havecker E.R., Gao X., Voytas D.F. The diversity of LTR retrotransposons // *Genome Biol.* 2004. Vol. 5, № 6. P. 225.
8. Malik H.S., Eickbush T.H. Phylogenetic Analysis of Ribonuclease H Domains Suggests a Late, Chimeric Origin of LTR Retrotransposable Elements and Retroviruses // *Genome Res.* 2001. Vol. 11, № 7. P. 1187–1197.
9. Ustyantsev K. et al. Convergent Evolution of Ribonuclease H in LTR Retrotransposons and Retroviruses // *Mol. Biol. Evol.* Oxford University Press, 2015. Vol. 32, № 5. P. 1197.
10. Hizi A., Herschhorn A. Retroviral reverse transcriptases (other than those of HIV-1 and murine leukemia virus): A comparison of their molecular and biochemical properties // *Virus Res.* 2008. Vol. 134, № 1. P. 203–220.
11. Bessereau J.-L. Transposons in *C. elegans* // *WormBook* / ed. The *C. elegans* Research Community. WormBook, 2006.
12. Kim Y.-J., Lee J., Han K. Transposable Elements: No More “Junk DNA”. // *Genomics Inform.* Korea Genome Organization, 2012. Vol. 10, № 4. P. 226–233.
13. Volff J.-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes // *BioEssays.* 2006. Vol. 28, № 9. P. 913–922.
14. Bennetzen J.L., Ma J., Devos K.M. Mechanisms of Recent Genome Size Variation in Flowering Plants // *Ann. Bot.* Oxford University Press, 2005. Vol. 95, № 1. P. 127–132.
15. Kumar A., Bennetzen J.L. Plant Retrotransposons // *Annu. Rev. Genet.* 1999. Vol. 33, № 1. P. 479–532.
16. Schnable P.S. et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics // *Science.* 2009. Vol. 326, № 5956. P. 1112–1115.

17. Chénais B. et al. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. // *Gene*. 2012. Vol. 509, № 1. P. 7–15.
18. SanMiguel P. et al. The paleontology of intergene retrotransposons of maize // *Nat. Genet.* 1998. Vol. 20, № 1. P. 43–45.
19. Sabot F., Schulman A.H. Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome // *Heredity*. 2006. Vol. 97, № 6. P. 381–388.
20. Deininger P. Induction of DNA rearrangement and transposition. // *Proc. Natl. Acad. Sci. U. S. A.* 1993. Vol. 90, № 9. P. 3780–3781.
21. Lister C., Jackson D., Martin C. Transposon-induced inversion in *Antirrhinum* modifies *nivea* gene expression to give a novel flower color pattern under the control of *cycloidearadialis*. // *Plant Cell*. 1993. Vol. 5, № 11. P. 1541–1553.
22. Sen S.K. et al. Human Genomic Deletions Mediated by Recombination between Alu Elements // *Am. J. Hum. Genet. Cell Press*, 2006. Vol. 79, № 1. P. 41–53.
23. Feschotte C. Transposable elements and the evolution of regulatory networks // *Nat. Rev. Genet.* 2008. Vol. 9, № 5. P. 397–405.
24. Hawkins J.S., Grover C.E., Wendel J.F. Repeated big bangs and the expanding universe: Directionality in plant genome size evolution // *Plant Sci.* 2008. Vol. 174, № 6. P. 557–562.
25. Sahebi M. et al. Contribution of transposable elements in the plant's genome // *Gene*. 2018. Vol. 665. P. 155–166.
26. Mita P., Boeke J.D. How retrotransposons shape genome regulation // *Curr. Opin. Genet. Dev.* 2016. Vol. 37. P. 90–100.
27. Barkan A., Martienssen R.A. Inactivation of maize transposon Mu suppresses a mutant phenotype by activating an outward-reading promoter near the end of Mu1. // *Proc. Natl. Acad. Sci. Proceedings of the National Academy of Sciences*, 1991. Vol. 88, № 8. P. 3502–3506.
28. Ito H. et al. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress: 7341 // *Nature*. Nature Publishing Group, 2011. Vol. 472, № 7341. P. 115–119.
29. Girard L., Freeling M. Mutator-suppressible alleles of *rough sheath1* and *liguleless3* in maize reveal multiple mechanisms for suppression // *Genetics*. 2000. Vol. 154, № 1. P. 437–446.
30. Chen W. et al. Transposon insertion in a cinnamyl alcohol dehydrogenase gene is responsible for a brown midrib1 mutation in maize // *Plant Mol. Biol.* 2012. Vol. 80, № 3. P. 289–297.
31. Luehrsen K.R., Walbot V. Insertion of Mu1 elements in the first intron of the *Adh1-S* gene of maize results in novel RNA processing events. // *Plant Cell*. 1990. Vol. 2, № 12. P. 1225–1238.
32. Papa C.M. et al. Maize Chromomethylase *Zea methyltransferase2* Is Required for CpNpG Methylation // *Plant Cell*. 2001. Vol. 13, № 8. P. 1919–1928.
33. Cowley M., Oakey R.J. Transposable elements re-wire and fine-tune the transcriptome // *PLoS Genet.* 2013. Vol. 9, № 1. P. e1003234.
34. Lisch D. How important are transposons for plant evolution? 1 // *Nat. Rev. Genet.* Nature Publishing Group, 2013. Vol. 14, № 1. P. 49–61.

35. Regulski M. et al. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA // *Genome Res.* 2013. Vol. 23, № 10. P. 1651–1662.
36. Ong-Abdullah M. et al. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm: 7570 // *Nature*. Nature Publishing Group, 2015. Vol. 525, № 7570. P. 533–537.
37. Zabala G., Vodkin L.O. Methylation Affects Transposition and Splicing of a Large CACTA Transposon from a MYB Transcription Factor Regulating Anthocyanin Synthase Genes in Soybean Seed Coats // *PLOS ONE*. Public Library of Science, 2014. Vol. 9, № 11. P. e111959.
38. Zhao D., Jiang N. Nested Insertions and Accumulation of Indels Are Negatively Correlated with Abundance of Mutator-Like Transposable Elements in Maize and Rice // *PLOS ONE*. Public Library of Science, 2014. Vol. 9, № 1. P. e87069.
39. Purugganan M., Wessler S. The splicing of transposable elements and its role in intron evolution // *Genetica*. 1992. Vol. 86, № 1. P. 295–303.
40. Fedoroff N.V. Transposable Elements, Epigenetics, and Genome Evolution // *Science*. American Association for the Advancement of Science, 2012. Vol. 338, № 6108. P. 758–767.
41. White S.E., Habera L.F., Wessler S.R. Retrotransposons in the flanking regions of normal plant genes: a role for copia-like elements in the evolution of gene structure and expression. // *Proc. Natl. Acad. Sci. Proceedings of the National Academy of Sciences*, 1994. Vol. 91, № 25. P. 11792–11796.
42. Ortiz D.F., Strommer J.N. The Mu1 maize transposable element induces tissue-specific aberrant splicing and polyadenylation in two *Adh1* mutants // *Mol. Cell. Biol.* American Society for Microbiology, 1990. Vol. 10, № 5. P. 2090–2095.
43. Flavell A.J., Pearce S.R., Kumar A. Plant transposable elements and the genome // *Curr. Opin. Genet. Dev.* 1994. Vol. 4, № 6. P. 838–844.
44. Bradley D. et al. Complementary floral homeotic phenotypes result from opposite orientations of a transposon at the *plena* locus of *antirrhinum* // *Cell*. 1993. Vol. 72, № 1. P. 85–95.
45. Le T.N. et al. Epigenetic regulation of intragenic transposable elements impacts gene transcription in *Arabidopsis thaliana* // *Nucleic Acids Res.* 2015. Vol. 43, № 8. P. 3911–3921.
46. Peaston A.E. et al. Retrotransposons Regulate Host Genes in Mouse Oocytes and Preimplantation Embryos // *Dev. Cell*. 2004. Vol. 7, № 4. P. 597–606.
47. Morton T. et al. Paired-End Analysis of Transcription Start Sites in *Arabidopsis* Reveals Plant-Specific Promoter Signatures // *Plant Cell*. 2014. Vol. 26, № 7. P. 2746–2760.
48. Mejía-Guerra M.K. et al. Core Promoter Plasticity Between Maize Tissues and Genotypes Contrasts with Predominance of Sharp Transcription Initiation Sites // *Plant Cell*. 2015. Vol. 27, № 12. P. 3309–3320.
49. Hirsch C.D., Springer N.M. Transposable element influences on gene expression in plants // *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* 2017. Vol. 1860, № 1. P. 157–165.
50. Zemach A. et al. The *Arabidopsis* Nucleosome Remodeler DDM1 Allows DNA

- Methyltransferases to Access H1-Containing Heterochromatin // *Cell*. 2013. Vol. 153, № 1. P. 193–205.
51. Li Q. et al. Genetic Perturbation of the Maize Methyome // *Plant Cell*. 2014. Vol. 26, № 12. P. 4602–4616.
 52. Eichten S.R. et al. Spreading of Heterochromatin Is Limited to Specific Families of Maize Retrotransposons // *PLOS Genet. Public Library of Science*, 2012. Vol. 8, № 12. P. e1003127.
 53. Butelli E. et al. Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges // *Plant Cell*. 2012. Vol. 24, № 3. P. 1242–1255.
 54. Settles A.M. et al. Duplication and Suppression of Chloroplast Protein Translocation Genes in Maize // *Genetics*. 2001. Vol. 157, № 1. P. 349–360.
 55. Hollister J.D., Gaut B.S. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression // *Genome Res*. 2009. Vol. 19, № 8. P. 1419–1428.
 56. Vicent C.M., Casacuberta J.M. Impact of transposable elements on polyploid plant genomes // *Ann. Bot.* 2017. Vol. 120, № 2. P. 195–207.
 57. Twell D. et al. Promoter analysis of genes that are coordinately expressed during pollen development reveals pollen-specific enhancer sequences and shared regulatory elements. // *Genes Dev*. 1991. Vol. 5, № 3. P. 496–507.
 58. Cavrak V.V. et al. How a Retrotransposon Exploits the Plant's Heat Stress Response for Its Activation // *PLOS Genet. Public Library of Science*, 2014. Vol. 10, № 1. P. e1004115.
 59. Jardim S.S. et al. Effects of heat and UV radiation on the mobilization of transposon mariner-Mos1 // *Cell Stress Chaperones*. Springer Netherlands, 2015. Vol. 20, № 5. P. 843–851.
 60. Jiang N. et al. An active DNA transposon family in rice: 6919 // *Nature*. Nature Publishing Group, 2003. Vol. 421, № 6919. P. 163–167.
 61. Kikuchi K. et al. The plant MITE mPing is mobilized in anther culture: 6919 // *Nature*. Nature Publishing Group, 2003. Vol. 421, № 6919. P. 167–170.
 62. Nakazaki T. et al. Mobilization of a transposon in the rice genome: 6919 // *Nature*. Nature Publishing Group, 2003. Vol. 421, № 6919. P. 170–172.
 63. Sun F.-J. et al. Common evolutionary trends for SINE RNA structures // *Trends Genet*. 2007. Vol. 23, № 1. P. 26–33.
 64. Wallau G.L., Ortiz M.F., Loreto E.L.S. Horizontal transposon transfer in eukarya: detection, bias, and perspectives. // *Genome Biol. Evol*. 2012. Vol. 4, № 8. P. 689–699.
 65. Diao X., Freeling M., Lisch D. Horizontal Transfer of a Plant Transposon // *PLOS Biol. Public Library of Science*, 2005. Vol. 4, № 1. P. e5.
 66. Civaň P., Švec M., Hauptvogel P. On the Coevolution of Transposable Elements and Plant Genomes // *J. Bot. Hindawi*, 2011. Vol. 2011. P. e893546.
 67. Haimlich S. et al. Widespread horizontal gene transfer between plants and their microbiota. *bioRxiv*, 2022. P. 2022.08.25.505314.
 68. Gao D. et al. Horizontal Transfer of Non-LTR Retrotransposons from Arthropods to Flowering Plants // *Mol. Biol. Evol*. 2018. Vol. 35, № 2. P. 354–364.

69. Wickell D.A., Li F.-W. On the evolutionary significance of horizontal gene transfers in plants // *New Phytol.* 2020. Vol. 225, № 1. P. 113–117.
70. Suh A. Horizontal Transfer of Transposons as Genomic Fossils of Host-Parasite Interactions // *The Evolution and Fossil Record of Parasitism: Coevolution and Paleoparasitological Techniques* / ed. De Baets K., Huntley J.W. Cham: Springer International Publishing, 2021. P. 451–463.
71. Levis R.W. et al. Transposons in place of telomeric repeats at a *Drosophila* telomere // *Cell.* 1993. Vol. 75, № 6. P. 1083–1093.
72. Eickbush T.H. Telomerase and retrotransposons: which came first? // *Science.* 1997. Vol. 277, № 5328. P. 911–912.
73. Cahill S. et al. Effect of BRAFV600E mutation on transcription and post-transcriptional regulation in a papillary thyroid carcinoma model // *Mol Cancer.* 2007. Vol. 6. P. 21.
74. Caronia L.M., Phay J.E., Shah M.H. Role of BRAF in thyroid oncogenesis // *Clin. Cancer Res.* 2011. Vol. 17, № 24. P. 7511–7517.
75. Schulten H.J. et al. Mutational screening of RET, HRAS, KRAS, NRAS, BRAF, AKT1, and CTNNB1 in medullary thyroid carcinoma // *Anticancer Res.* 2011. Vol. 31, № 12. P. 4179–4183.
76. Bureau T.E., White S.E., Wessler S.R. Transduction of a cellular gene by a plant retroelement // *Cell.* 1994. Vol. 77, № 4. P. 479–480.
77. Cordaux R. et al. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element // *Proc. Natl. Acad. Sci. U. S. A.* 2006. Vol. 103, № 21. P. 8101–8106.
78. Blond J.L. et al. An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. // *J. Virol. American Society for Microbiology (ASM),* 2000. Vol. 74, № 7. P. 3321–3329.
79. Evgen'ev M.B., Arkhipova I.R. Penelope-like elements – a new class of retroelements: distribution, function and possible evolutionary significance // *Cytogenet. Genome Res. Karger Publishers,* 2005. Vol. 110, № 1–4. P. 510–521.
80. Arkhipova I.R. Distribution and Phylogeny of Penelope-Like Elements in Eukaryotes // *Syst. Biol.* 2006. Vol. 55, № 6. P. 875–885.
81. Evgen'ev M.B. et al. Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in *Drosophila virilis* // *Proc. Natl. Acad. Sci. Proceedings of the National Academy of Sciences,* 1997. Vol. 94, № 1. P. 196–201.
82. Finnegan D.J. Retrotransposons // *Curr. Biol.* 2012. Vol. 22, № 11. P. R432–R437.
83. Kidwell M.G. Transposable elements and the evolution of genome size in eukaryotes // *Genetica. Kluwer Academic Publishers,* 2002. Vol. 115, № 1. P. 49–63.
84. Eickbush T.H., Malik H.S. Origins and Evolution of Retrotransposons // *Mob. DNA II.* 2002. P. 1111–1144.
85. Jin Y.K., Bennetzen J.L. Structure and coding properties of Bsl, a maize retrovirus-like transposon. // *Proc. Natl. Acad. Sci. U. S. A.* 1989. Vol. 86, № 16. P. 6235–6239.

86. Hull R., Covey S.N. Retroelements: Propagation and adaptation // *Virus Genes*. 1995. Vol. 11, № 2. P. 105–118.
87. Malik H.S., Eickbush T.H. Modular Evolution of the Integrase Domain in the Ty3/Gypsy Class of LTR Retrotransposons // *J. Virol.* 1999. Vol. 73, № 6. P. 5186–5190.
88. Rausch J.W., Miller J.T., Le Grice S.F.J. Reverse Transcription in the *Saccharomyces cerevisiae* Long-Terminal Repeat Retrotransposon Ty3 // *Viruses*. 2017. Vol. 9, № 3. P. 44.
89. Novikova O. et al. Novel clades of chromodomain-containing Gypsy LTR retrotransposons from mosses (Bryophyta) // *Plant J.* 2008. Vol. 56, № 4. P. 562–574.
90. Leblanc P. et al. Life Cycle of an Endogenous Retrovirus, ZAM, in *Drosophila melanogaster* // *J. Virol.* 2000. Vol. 74, № 22. P. 10658–10669.
91. Menéndez-Arias L., Sebastián-Martín A., Álvarez M. Viral reverse transcriptases // *Virus Res.* 2017. Vol. 234. P. 153–176.
92. Retrotransposon // Wikipedia. 2019.
93. Herschhorn A., Hizi A. Retroviral reverse transcriptases // *Cell. Mol. Life Sci.* 2010. Vol. 67, № 16. P. 2717–2747.
94. Figiel M. et al. Mechanism of polypurine tract primer generation by HIV-1 reverse transcriptase // *J. Biol. Chem.* 2018. Vol. 293, № 1. P. 191–202.
95. Schulman A.H. Hitching a Ride: Nonautonomous Retrotransposons and Parasitism as a Lifestyle // *Plant Transposable Elements: Impact on Genome Structure and Function* / ed. Grandbastien M.-A., Casacuberta J.M. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. P. 71–88.
96. Muşat M.G. et al. HIV-1 integrase inhibitors targeting various DDE transposases: Retroviral integration versus RAG-mediated recombination (Review) // *Mol. Med. Rep. Spandidos Publications*, 2019. Vol. 20, № 6. P. 4749–4762.
97. Rice P.A., Baker T.A. Comparative architecture of transposase and integrase complexes: 4 // *Nat. Struct. Biol. Nature Publishing Group*, 2001. Vol. 8, № 4. P. 302–307.
98. Cerbin S. et al. GingerRoot: A Novel DNA Transposon Encoding Integrase-Related Transposase in Plants and Animals // *Genome Biol. Evol.* 2019. Vol. 11, № 11. P. 3181–3193.
99. Bao W., Kapitonov V.V., Jurka J. Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons // *Mob. DNA*. 2010. Vol. 1, № 1. P. 3.
100. Feschotte C., Pritham E.J. DNA Transposons and the Evolution of Eukaryotic Genomes // *Annu. Rev. Genet.* 2007. Vol. 41, № 1. P. 331–368.
101. Ebina H. et al. The GP(Y/F) Domain of TF1 Integrase Multimerizes when Present in a Fragment, and Substitutions in This Domain Reduce Enzymatic Activity of the Full-length Protein // *J. Biol. Chem.* 2008. Vol. 283, № 23. P. 15965–15974.
102. Maertens G.N., Engelman A.N., Cherepanov P. Structure and function of retroviral integrase: 1 // *Nat. Rev. Microbiol. Nature Publishing Group*, 2022. Vol. 20, № 1. P. 20–34.
103. Li F. et al. Identification of an integrase-independent pathway of

- retrotransposition // *Sci. Adv. American Association for the Advancement of Science*, 2022. Vol. 8, № 26. P. eabm9390.
104. Rai S.K. et al. Host factors that promote retrotransposon integration are similar in distantly related eukaryotes // *PLoS Genet.* 2017. Vol. 13, № 12. P. e1006775.
 105. Engelman A. et al. Multiple effects of mutations in human immunodeficiency virus type 1 integrase on viral replication // *J. Virol.* 1995. Vol. 69, № 5. P. 2729–2736.
 106. Leavitt A.D. et al. Human immunodeficiency virus type 1 integrase mutants retain in vitro integrase activity yet fail to integrate viral DNA efficiently during infection // *J. Virol.* 1996. Vol. 70, № 2. P. 721–728.
 107. Jasin M., Rothstein R. Repair of strand breaks by homologous recombination // *Cold Spring Harb. Perspect. Biol.* 2013. Vol. 5, № 11. P. a012740.
 108. Bai Y., Davis A.P., Symington L.S. A novel allele of RAD52 that causes severe DNA repair and recombination deficiencies only in the absence of RAD51 or RAD59 // *Genetics*. 1999. Vol. 153, № 3. P. 1117–1130.
 109. Shi I. et al. Role of the Rad52 amino-terminal DNA binding activity in DNA strand capture in homologous recombination // *J. Biol. Chem.* 2009. Vol. 284, № 48. P. 33275–33284.
 110. Hyjek M., Figiel M., Nowotny M. RNases H: Structure and mechanism // *DNA Repair*. 2019.
 111. Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons // *Proc. Natl. Acad. Sci.* 1997. Vol. 94, № 5. P. 1872–1877.
 112. Nowotny M. Retroviral integrase superfamily: the structural perspective // *EMBO Rep.* 2009. Vol. 10, № 2. P. 144–151.
 113. Engelman A., Mizuuchi K., Craigie R. HIV-1 DNA integration: Mechanism of viral DNA cleavage and DNA strand transfer // *Cell. Cell Press*, 1991. Vol. 67, № 6. P. 1211–1221.
 114. Witt R.L. et al. Diagnosis and management of differentiated thyroid cancer using molecular biology. // *The Laryngoscope*. 2013. Vol. 123, № 4. P. 1059–1064.
 115. Malik H.S. Ribonuclease H evolution in retrotransposable elements // *Cytogenet. Genome Res. Karger Publishers*, 2005. Vol. 110, № 1–4. P. 392–401.
 116. Cerritelli S.M., Crouch R.J. Ribonuclease H: the enzymes in Eukaryotes // *FEBS J.* 2009. Vol. 276, № 6. P. 1494.
 117. Figiel M. et al. Coordination between the polymerase and RNase H activity of HIV-1 reverse transcriptase // *Nucleic Acids Res.* 2017. Vol. 45, № 6. P. 3341–3352.
 118. Ustyantsev K., Blinov A., Smyshlyaev G. Convergence of retrotransposons in oomycetes and plants // *Mob. DNA. BioMed Central*, 2017. Vol. 8, № 1. P. 4.
 119. Steinbauerová V. et al. A widespread occurrence of extra open reading frames in plant Ty3/gypsy retrotransposons // *Genetica*. 2011. Vol. 139, № 11. P. 1543–1555.
 120. Smyshlyaev G. et al. Acquisition of an Archaea-like ribonuclease H domain by plant L1 retrotransposons supports modular evolution. // *Proc. Natl. Acad. Sci. U. S. A.* 2013. Vol. 110, № 50. P. 20140–20145.
 121. Abrusán G. et al. Turning gold into ‘junk’: transposable elements utilize central

- proteins of cellular networks // *Nucleic Acids Res.* 2013. Vol. 41, № 5. P. 3190–3200.
122. Nowak E. et al. Ty3 reverse transcriptase complexed with an RNA-DNA hybrid shows structural and functional asymmetry. // *Nat. Struct. Mol. Biol.* 2014. Vol. 21, № 4. P. 389–396.
 123. Ellinghaus D., Kurtz S., Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons // *BMC Bioinformatics.* 2008. Vol. 9, № 1. P. 18.
 124. Sanchez D.H. et al. High-frequency recombination between members of an LTR retrotransposon family during transposition bursts: 1 // *Nat. Commun. Nature Publishing Group*, 2017. Vol. 8, № 1. P. 1283.
 125. Cock P.J.A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics // *Bioinformatics. Oxford Academic*, 2009. Vol. 25, № 11. P. 1422–1423.
 126. Camacho C. et al. BLAST+: architecture and applications // *BMC Bioinformatics.* 2009. Vol. 10, № 1. P. 421.
 127. Marchler-Bauer A. et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures // *Nucleic Acids Res.* 2017. Vol. 45, № Database issue. P. D200–D203.
 128. Fu L. et al. CD-HIT: accelerated for clustering the next-generation sequencing data // *Bioinforma. Oxf. Engl.* 2012. Vol. 28, № 23. P. 3150–3152.
 129. Steinegger M., Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets // *Nat. Biotechnol.* 2017. Vol. 35, № 11. P. 1026–1028.
 130. Katoh K., Standley D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. // *Mol. Biol. Evol.* 2013. Vol. 30, № 4. P. 772–780.
 131. Katoh K., Rozewicki J., Yamada K.D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization // *Brief. Bioinform.* 2017.
 132. Llorens C. et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0 // *Nucleic Acids Res.* 2011. Vol. 39, № suppl_1. P. D70–D74.
 133. Price M.N., Dehal P.S., Arkin A.P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments // *PLoS ONE* / ed. Poon A.F.Y. 2010. Vol. 5, № 3. P. e9490.
 134. Nguyen L.-T. et al. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies // *Mol. Biol. Evol. Oxford University Press*, 2015. Vol. 32, № 1. P. 268–274.
 135. Trifinopoulos J. et al. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis // *Nucleic Acids Res.* 2016. Vol. 44, № W1. P. W232–W235.
 136. Anisimova M., Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative // *Syst. Biol.* 2006. Vol. 55, № 4. P. 539–552.
 137. UFBoot2: Improving the Ultrafast Bootstrap Approximation | *Molecular*

- Biology and Evolution | Oxford Academic [Electronic resource]. URL: <https://academic.oup.com/mbe/article/35/2/518/4565479> (accessed: 16.09.2021).
138. Berman H.M. et al. The Protein Data Bank // *Nucleic Acids Res.* 2000. Vol. 28, № 1. P. 235–242.
 139. Pei J., Kim B.-H., Grishin N.V. PROMALS3D: a tool for multiple protein sequence and structure alignments // *Nucleic Acids Res.* 2008. Vol. 36, № 7. P. 2295–2300.
 140. Vicient C.M., Casacuberta J.M. Additional ORFs in Plant LTR-Retrotransposons // *Front. Plant Sci.* 2020. Vol. 11.
 141. Home - ORFfinder - NCBI [Electronic resource]. URL: <https://www.ncbi.nlm.nih.gov/orffinder/> (accessed: 14.09.2021).
 142. Biryukov M., Ustyantsev K. DARTS: An Algorithm for Domain-Associated Retrotransposon Search in Genome Assemblies: 1 // *Genes. Multidisciplinary Digital Publishing Institute*, 2022. Vol. 13, № 1. P. 9.
 143. Leebens-Mack J.H. et al. One thousand plant transcriptomes and the phylogenomics of green plants: 7780 // *Nature*. Nature Publishing Group, 2019. Vol. 574, № 7780. P. 679–685.
 144. Wang X.-Q., Ran J.-H. Evolution and biogeography of gymnosperms // *Mol. Phylogenet. Evol.* 2014. Vol. 75. P. 24–40.
 145. Sessa E.B. et al. Between Two Fern Genomes // *GigaScience*. 2014. Vol. 3, № 1. P. 2047-217X-3–15.
 146. Uddenberg D. et al. Sequenced genomes and rapidly emerging technologies pave the way for conifer evolutionary developmental biology // *Front. Plant Sci.* 2015. Vol. 6. P. 970.
 147. Lu S. et al. CDD/SPARCLE: the conserved domain database in 2020 // *Nucleic Acids Res.* 2020. Vol. 48, № D1. P. D265–D268.
 148. Wang J., Han G.-Z. A Sister Lineage of Sampled Retroviruses Corroborates the Complex Evolution of Retroviruses // *Mol. Biol. Evol.* 2021. Vol. 38, № 3. P. 1031–1039.
 149. Wang J., Han G.-Z. A Missing Link between Retrotransposons and Retroviruses // *mBio. American Society for Microbiology*, 2022. Vol. 13, № 2. P. e00187-22.
 150. Lapkouski M. et al. Complexes of HIV-1 RT, NNRTI and RNA/DNA hybrid reveal a structure compatible with RNA degradation // *Nat. Struct. Mol. Biol.* 2013. Vol. 20, № 2. P. 230–236.
 151. Chung S. et al. Examining the Role of the HIV-1 Reverse Transcriptase p51 Subunit in Positioning and Hydrolysis of RNA/DNA Hybrids // *J. Biol. Chem.* 2013. Vol. 288, № 22. P. 16177–16184.
 152. Das K., Sarafianos S.G., Arnold E. Structural requirements for RNA degradation by HIV-1 reverse transcriptase // *Nat. Struct. Mol. Biol.* 2013. Vol. 20, № 12. P. 1341–1342.
 153. Zennou V. et al. Loss of Viral Fitness Associated with Multiple Gag and Gag-Pol Processing Defects in Human Immunodeficiency Virus Type 1 Variants Selected for Resistance to Protease Inhibitors In Vivo // *J. Virol. American Society for Microbiology*, 1998. Vol. 72, № 4. P. 3300–3306.
 154. Lin Y.-R. et al. Effects of reduced gag cleavage efficiency on HIV-1 Gag-Pol

- package // BMC Microbiol. 2022. Vol. 22, № 1. P. 94.
155. Könnyű B. et al. Gag-Pol Processing during HIV-1 Virion Maturation: A Systems Biology Approach // PLOS Comput. Biol. Public Library of Science, 2013. Vol. 9, № 6. P. e1003103.
156. Shehu-Xhilaga M., Crowe S.M., Mak J. Maintenance of the Gag/Gag-Pol Ratio Is Important for Human Immunodeficiency Virus Type 1 RNA Dimerization and Viral Infectivity // J. Virol. American Society for Microbiology, 2001. Vol. 75, № 4. P. 1834–1841.
157. Jiao Y., Guo H. Chapter Nine - Prehistory of the Angiosperms: Characterization of the Ancient Genomes // Advances in Botanical Research / ed. Paterson A.H. Academic Press, 2014. Vol. 69. P. 223–245.

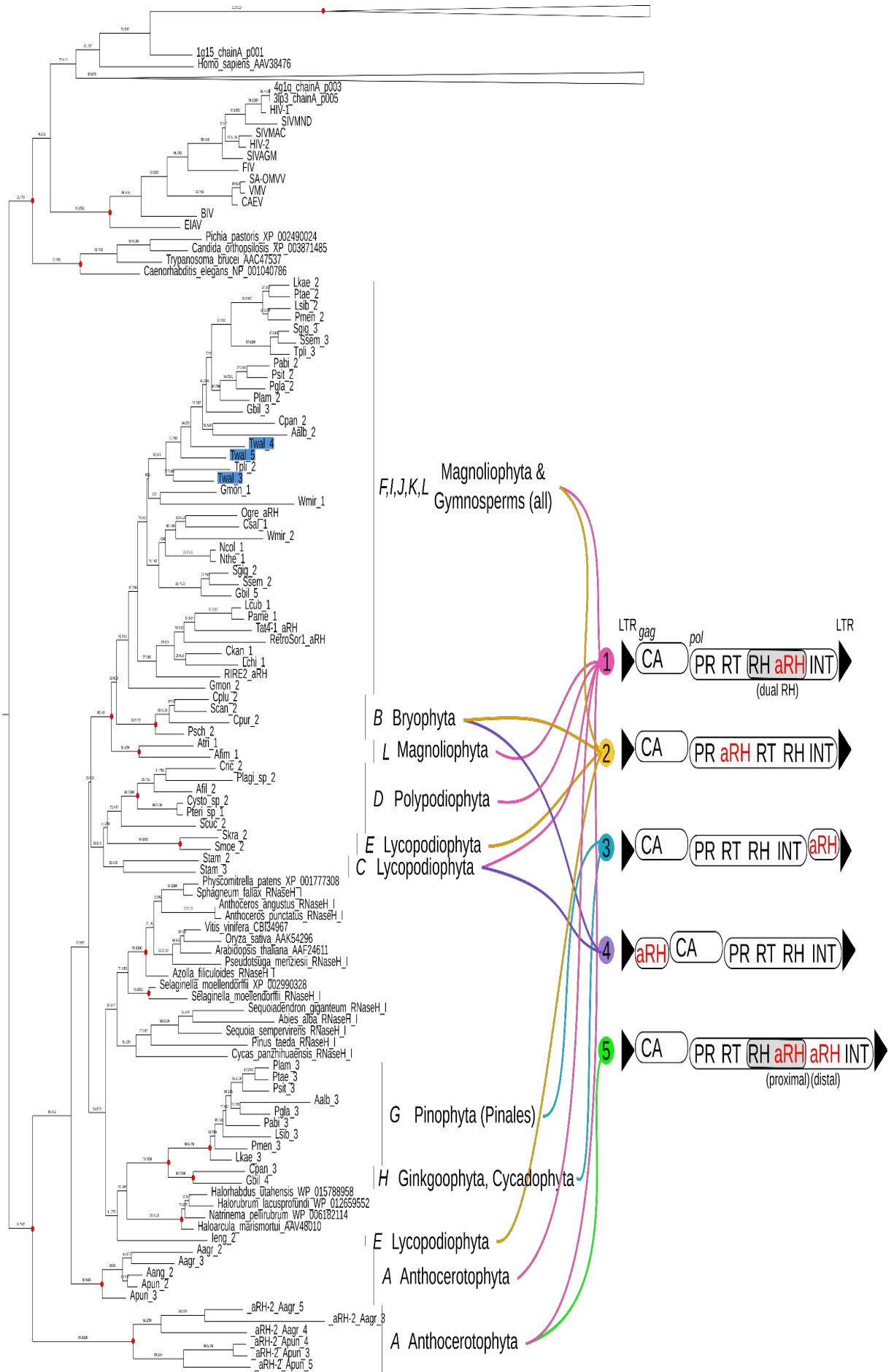
Приложения

Приложение 1. Таблица геномов зелёных растений, исследованных в данной работе (за исключением зеленых водорослей) и их таксономической принадлежности и указания количества элементов *Tat* с доменов aRNH, обнаруженных в этих геномах. Структуры по положению aRNH схематично отображены на рисунке 10. Деревья, по которым можно обнаружить данных представителей представлены в приложениях 2 и 4. Таксоны водорослей, исследованные в рамках данной работы не приведены в силу высокой численности геномов и отсутствия в каждом из них элементов с aRNH.

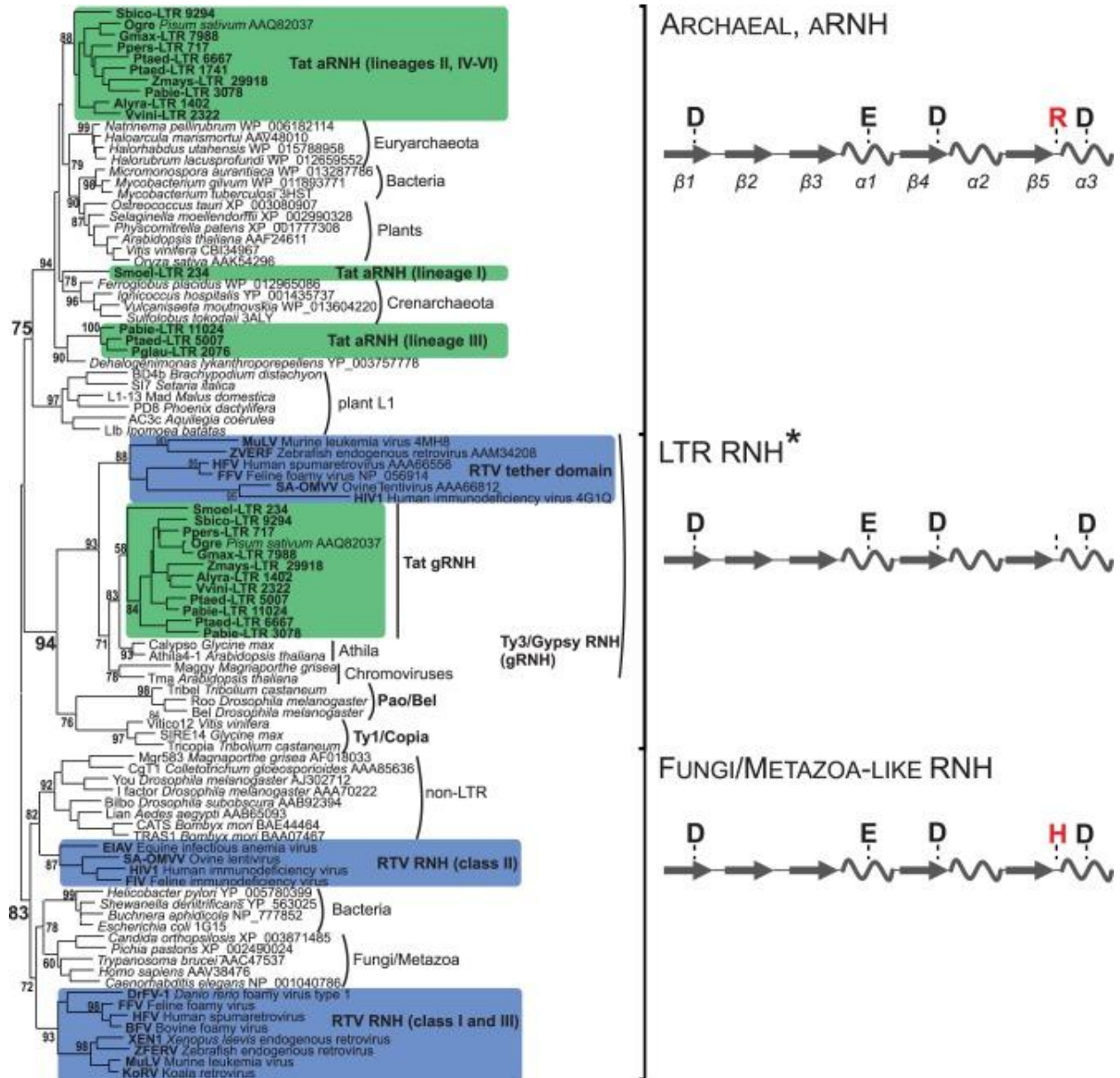
Основная группа	Класс	Вид	Количество	Структуры
Клада Marchantiophyta	Marchantiopsida	<i>Marchantia polymorpha</i>	0	-
		<i>Marchantia inflexa</i>	0	-
Клада Anthocerotophyta	Anthocerotopsida	<i>Anthoceros angustus</i>	4	#1
		<i>Anthoceros punctatus</i>	9	#1,#5
		<i>Anthoceros agrestis</i>	19	#1,#5
Клада Bryophyta	Sphagnopsida	<i>Sphagnum magellanicum</i>	0	-
		<i>Sphagnum fallax</i>	0	-
	Bryopsida	<i>Physcomitrella patens</i>	0	-
		<i>Pleurozium schreberi</i>	7	#4
		<i>Calohypnum plumiforme</i>	8	#4
		<i>Ceratodon purpureus</i>	5	#4
		<i>Fontinalis antipyretica</i>	0	-
		<i>Syntrichia caninervis</i>	12	#4
Класс Lycopodiopsida	Lycopodiopsida	<i>Selaginella tamariscina</i>	73	#1,#4
		<i>Selaginella moellendorffii</i>	51	#2
		<i>Selaginella kraussiana</i>	2	#2,#3
		<i>Isoetes engelmannii</i>	110	#2
Класс Polypodiopsida	Polypodiopsida	<i>Polypodium sp</i>	0	-
		<i>Pteridium sp</i>	3	#1
		<i>Plagiogyria sp</i>	9	#1
		<i>Dipteris sp</i>	0	-
		<i>Cystopteris sp</i>	4	#1
		<i>Ceratopteris sp</i>	0	-
		<i>Ceratopteris richardii</i>	53	#1
		<i>Salvinia cucullata</i>	17	#1
		<i>Azolla filiculoides</i>	167	#1

Клада Gnetophyta	Gnetopsida	<i>Gnetum montanum</i>	26160	#1,#2
		<i>Welwitschia mirabilis</i>	879	#1
Клада Ginkgophyta	Ginkgoopsida	<i>Ginkgo biloba</i>	21357	#1,#2,#3
Клада Cycadophyta	Cycadopsida	<i>Cycas panzhihuaensis</i>	2922	#2,#3
Класс Pinopsida	Pinopsida (conifers I clade - отдел Pinales, хвойные)	<i>Pseudotsuga menziesii</i>	65496	#2,#3
		<i>Abies alba</i>	86303	#2,#3
		<i>Pinus taeda</i>	39538	#2,#3
		<i>Pinus lambertiana</i>	49880	#2,#3
		<i>Picea abies</i>	29708	#2,#3
		<i>Larix sibirica</i>	2135	#2,#3
		<i>Picea sitchensis</i>	50267	#2,#3
		<i>Larix kaempferi</i>	27181	#2,#3
		<i>Picea glauca</i>	21831	#2,#3
	Pinopsida (conifers II clade - отдел Cupressales, кипарисовые и тисовые)	<i>Sequoiadendron giganteum</i>	56032	#1,#2
		<i>Sequoia sempervirens</i>	164834	#1,#2
		<i>Thuja plicata</i>	49262	#1,#2
		<i>Taxus wallichiana</i>	73202	#1,#2
Класс Magnoliopsida	Magnoliopsida	<i>Amborella trichopoda</i>	1688	#1
		<i>Nymphaea colorata</i>	314	#1
		<i>Nymphaea thermarum</i>	158	#1
		<i>Aristolochia fimbriata</i>	1302	#1
		<i>Liriodendron chinense</i>	12825	#1
		<i>Chimonanthus salicifolius</i>	1	#1
		<i>Cinnamomum kanehirae</i>	1587	#1
		<i>Persea americana</i>	2804	#1
		<i>Litsea cubeba</i>	10664	#1

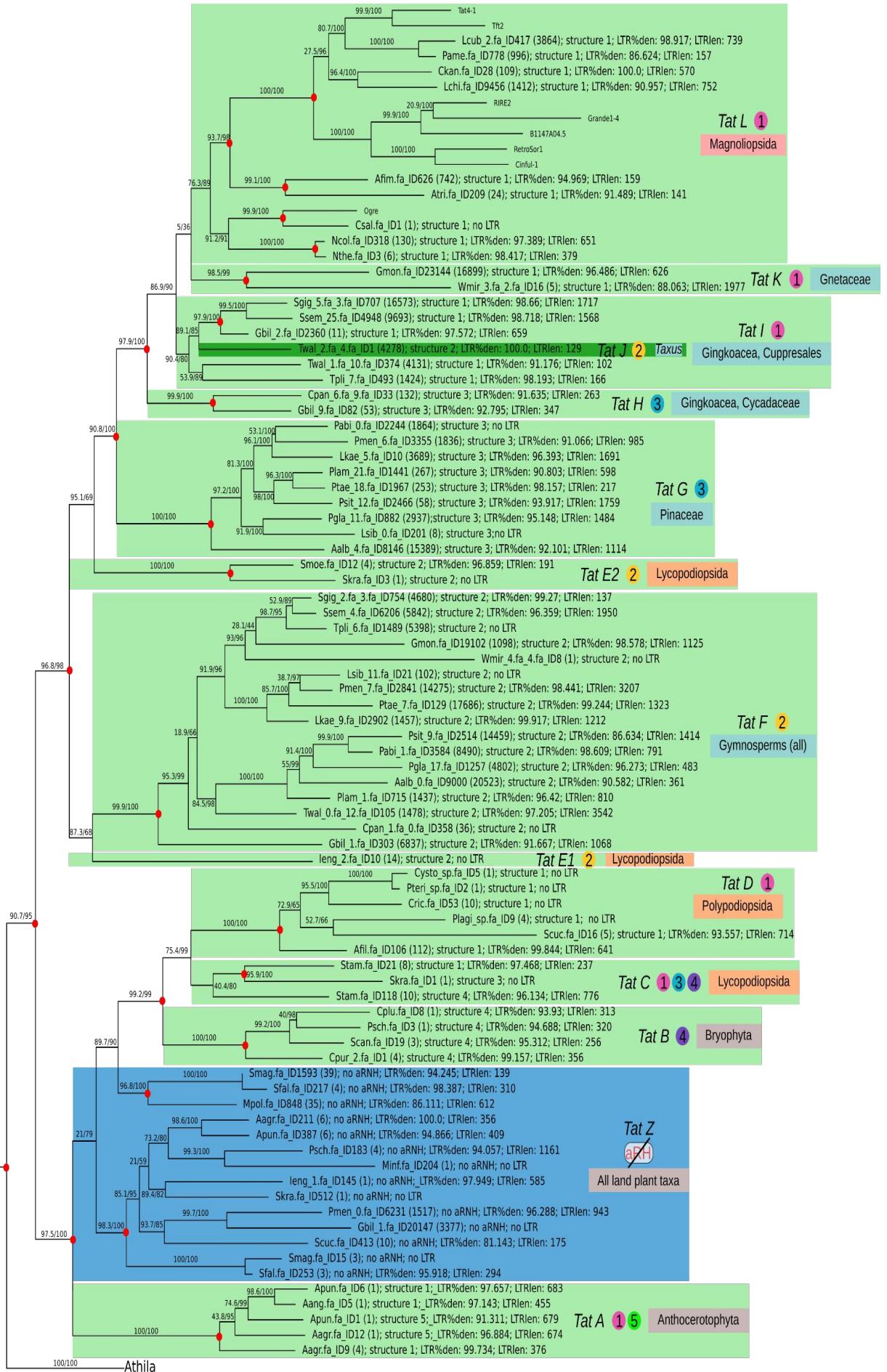
Приложение 2. Филогенетическое древо домена aRNH в сопоставлении со структурами, построенное на элементах зелёных растений. Сверху упрощены все домены “LTR RNH”, разделяющиеся далее на RNH *Ty3/Gypsy*, *Ty1/Copia* и *Bel/Pao*. Вместе с ними в качестве ещё одной внешней группы взяты домены RNH ретровирусов. Вся нижняя ветвь - домены aRNH. Подписи в этом блоке соответствуют новой нумерации *Tat* (рисунок 11). Внутри ветви “aRNH *Tat F,I,J,K,L*” три последовательности генома *Taxus wallichiana* из трёх кластеров *F,I,J* выделены синим в следующем порядке: верхняя соответствует элементам кластера *Tat F* со структурой #2, нижняя - элементам кластера *I* со структурой #1, а средняя - #2 из кластера *Tat J*. Достоверные по коэффициентам поддержки (aLRT и ufBoot) узлы выделены красным. Подписи элементов упрощены во избежание перегрузки изображения информацией.



Приложение 3. Филогенетическое древо домена aRNH, адаптировано из прежней статьи [9], построенное преимущественно на элементах покрытосеменных. Как и в аналогичном древе, выполненном в рамках этой работы (схематично изображённом на рисунке 12, в полном виде представлено в приложении 2), выделяется три основных больших кластера доменов RNH: (1) fmRNH, или свойственные грибам и животным (Fungi/Metazoa-like RNH), (2) нативные RNH LTR-ретротранспозонов (LTR RNH), (3) дополнительный домен aRNH, называемый в прежних исследованиях “архейным” (Archeal RNH) - от организмов, чьи собственные домены RNH также попали в этот кластер и которые рассматривались как вероятный источник приобретения элементами *Tat* домена aRNH. Для каждого из трёх основных кластеров показана третичная структура (α -спирали и β -складки) с 4(5) аминокислотами каталитически активного центра DED(R/H/-)D.



Приложение 4. Древо филогенетических взаимоотношений LTR-ретротранспозонов, построенное на основании выравнивания аминокислотных последовательностей домена RT из данной работы. В качестве внешних групп использовались элементы *Ty3/Gypsy* (не отражены на древе) и наиболее близкородственной к элементам *Tat* группе - *Athila* (схлопнуто внизу древа в единую линию). и сам *Tat*, разбитый на 13 внутренних кластеров *A-L* и *Z* ("Zero"), не содержащий добавочного домена aRNH. Выявленные элементы подписаны в формате "количество элементов в кластере, сокращение видового названия, файл после разбиения исходного файла генома на файлы меньшего размера, структура по доменам, характеристики LTR при наличии, счётчик-показатель качества элементов". Кластер *Tat E* разбит на две части из-за неустойчивого положения элементов генома плауна *Isoetes engelmannii*. Кластер *Tat J*, представленный элементами одного генома из рода *Taxus*, отражает повторное приобретение структуры #2 среди элементов со структурой #1 (*Tat I*). Достоверные по коэффициентам поддержки (aLRT и ufBoot) узлы выделены красным.



Приложение 5. Таблица элементов, найденных в геномах после работа алгоритма DARTS, определённые запросом в командной строке через функцию 'grep'. Численности элементов могут отличаться от данных в таблице приложения 1, поскольку здесь учитываются химерные элементы и свободные домены aRNH, распознанные как полноценные элементы из-за наличия в непосредственной близости фрагментов RT, которые были отсеяны фильтром по длине при аннотации последовательности.

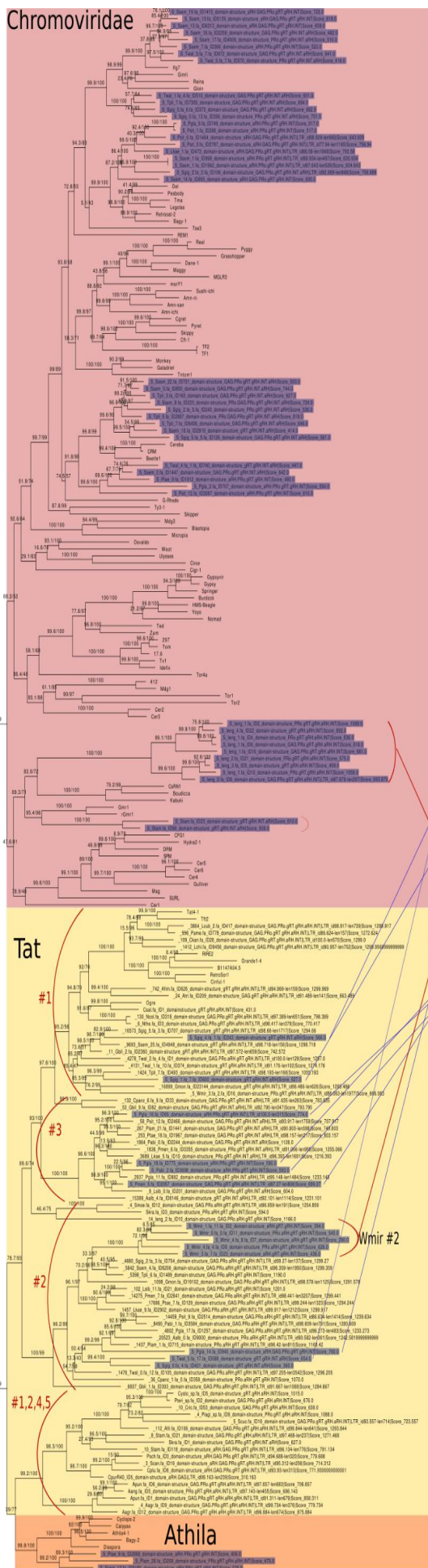
Вид	Состав по структурам						Одиночные aRNH	Комментарии
	#1	#2	#3	#4	#5	всего		
<i>M. polymorpha</i>	0					0	0	
<i>M. inflexa</i>	0					0	0	
<i>A. angustus</i>	4					4	0	
<i>A. punctatus</i>	6				3	9	0	
<i>A. agrestis</i>	16				3	19	0	
<i>S. magellanicum</i>	0					0	0	
<i>S. fallax</i>	0					0	0	
<i>P. patens</i>	0					0	0	
<i>P. schreberi</i>				7		7	0	
<i>C. plumiforme</i>				8		8	0	
<i>C. purpureus</i>				5		5	0	
<i>F. antipyretica</i>						0	0	
<i>S. caninervis</i>				11		12	0	2 химеры (1 ~#4)
<i>S. tamariscina</i>	15		2	51		73	1	
<i>S. moellendorffii</i>		51				51	0	
<i>S. kraussiana</i>		1	1			2	0	
<i>I. engelmannii</i>	13	96				110	1	
<i>Polypodium sp</i>	0					0	0	
<i>Pteridium sp</i>	3					3	0	
<i>Plagiogyria sp</i>	9					9	0	
<i>Dipteris sp</i>	0					0	0	
<i>Cystopteris sp</i>	4					4	0	

<i>Ceratopteris sp</i>	0					0	0	
<i>C. richardii</i>	53					53	0	
<i>S. cucullata</i>	17					17	0	
<i>A. filiculoides</i>	167					167	0	
<i>G. montanum</i>	23178	2903				26150	6	
<i>W. mirabilis</i>	791	5				879	0	
<i>G. biloba</i>	4969	15125	809			21357	213	35 ~#4
<i>C. panzhihuaensis</i>		177	1123			2906	1255	137 ~#1 без INT и 172 без INT+gRH
<i>P. menziesii</i>	1	43303	20664			65496	65	3 ~#1 без gRH, 296 без INT
<i>A. alba</i>		63113	21691			86276	74	
<i>P. taeda</i>		30971	7385	2		39538	67	917 ~#1 без INT
<i>P. lambertiana</i>	1	41314	7422			49880	65	729 ~#1 без INT
<i>P. abies</i>		25632	3709	1		29708	18	337 химеры из #1,#2,#3
<i>L. sibirica</i>		1969	33			2135	22	
<i>P. sitchensis</i>		43033	6782	5		50267	75	249 ~#1 без INT
<i>L. kaempferi</i>		13605	12794	1		27181	66	660 ~#1 без INT или INT+gRH
<i>P. glauca</i>		17494	3891	5		21831	46	384 ~#1 без INT или химеры
<i>S. giganteum</i>	36387	19459	9			56032	25	
<i>S. sempervirens</i>	97769	66249	7	8		164834	68	
<i>T. plicata</i>	31128	17884	4			49262	25	
<i>T. wallichiana</i>	25090	47335	5			73178	58	
<i>A. trichopoda</i>	1698					1702	0	

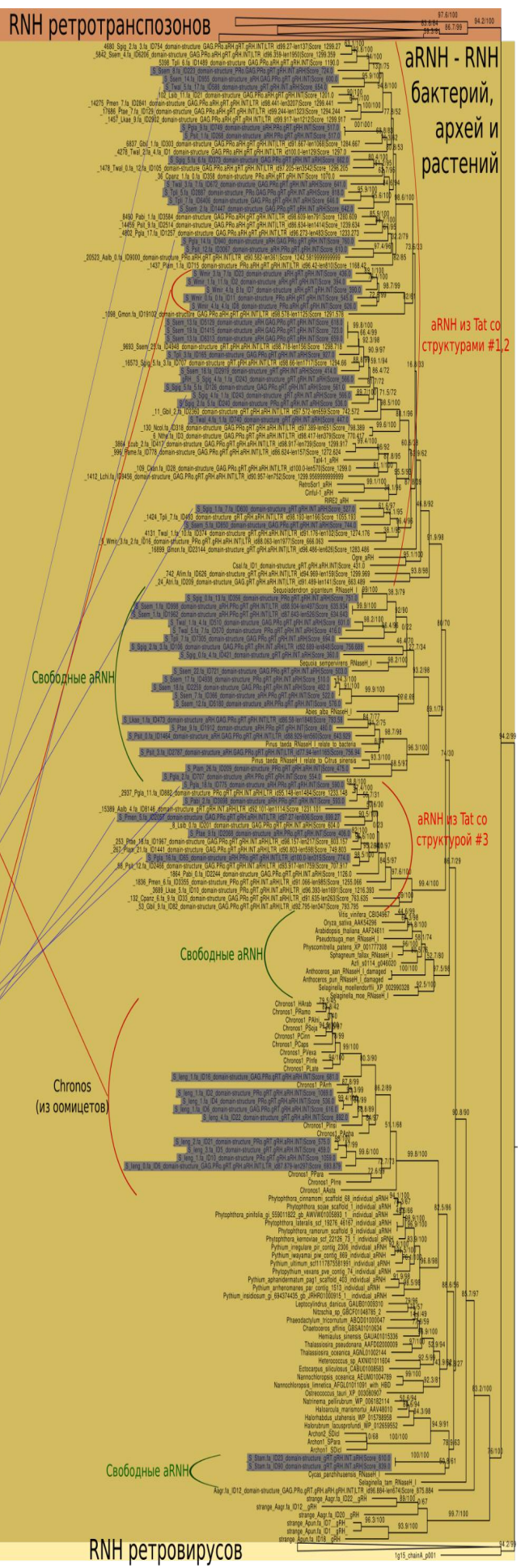
<i>N. colorata</i>	314					314	0	
<i>N. thermarum</i>	160					160	0	
<i>A. fimbriata</i>	1320					1322	1	
<i>L. chinense</i>	12978					13029	3	1 ~#3
<i>C. salicifolius</i>	1					1	0	
<i>C. kanehirae</i>	1607					1624	1	
<i>P. americana</i>	2846					2862	2	
<i>L. cubeba</i>	10736					10800	1	

Приложение 6. Сравнение филогенетических деревьев с добавлением синглетных последовательностей. Основные кандидаты в новые кластеры среди элементов-синглетов выделены синим фоном, среди них были следующие представители: элементы плауна *Isoetes engelmannii* со структурой #1, плауна *Selaginella tamariscina* со структурой #3, элементы гнетового *Welwitschia mirabilis* со структурой #2, элементы хвойных со структурами #1,4 и кипарисовых с #3,4. На деревьях не присутствует полный набор выявленных элементов - сокращены кластеры, в которые не попал ни один из исследуемых синглетов. В силу соответствия кластеру *Tat* на древе RT и кластерам aRNH из элементов *Tat* на древе RNH, элементы, соединённые синими линиями, могут быть как артефактом метода (химерные структуры деградировавших элементов распознались за полноценный элемент), так и истинными элементами со структурами, не выявленными систематически. В прочих случаях элементы либо не имели соответствия на деревьях - большая часть синглетов голосеменных либо не принадлежали кластеру *Tat* по домену RT, либо имели aRNH, соответствующий aRNH хозяйского генома или его свободным копиям (которые, вероятно, остались от давно деградировавших элементов). Элементы плауна *Isoetes engelmannii*, предположительно, являются контаминацией геномной сборки (собрана до контингов, не хромосом) паразитами растений - оомицетами (подписано - “*Chronos*”). Элементы *Welwitschia mirabilis* расположились в ожидаемом кластере на каждом из филогенетических деревьях (подписано “Wmir #2”).

Chromoviridae



RNH ретротранспозонов



Приложение 7. Информация о количестве структур, кластеров и характеристиках кластеров для всех выявленных элементов *Tat* в исследованных геномах. Для каждого сочетания геном-структура представлено количество полученных элементов с и без aRNH; структуры и доля каждой от общего количества элементов с aRNH; количество линий после кластеризации; доля элементов с обнаруженными последовательностями LTR; средней и максимальной идентичностью LTR среди этих элементов; количество элементов со 100% идентичными LTR.

Вид	Кол-во элементов в Tat без aRNH (Tat Z), N	Кол-во элементов Tat с aRNH, N	Структура, # и её доля среди всех элементов с aRNH (%)*	Кол-во линий после кластеризации, N	Элементы с обнаруженными LTRs, %	Средняя идентичность LTR, %	Макс. идентичность LTR, %	Кол-во элементов со 100% LTR, N
<i>M. polymorpha</i>	178	0	-	-	-	-	-	-
<i>M. inflexa</i>	148	0	-	-	-	-	-	-
<i>A. angustus</i>	11	4	#1 (100)	Синглеты	25.00	97.1	97.1	0
<i>A. punctatus</i>	32	9	#1 (66.7)	Синглеты	33.33	91.7	97.7	0
			#5 (33.3)	Синглеты	33.33	91.3	91.3	0
<i>A. agrestis</i>	72	19	#1 (84.2)	2	50.0	99.7	99.7	0
			#5 (15.8)	Синглеты	66.67	96.5	96.9	0
<i>S. magellanicum</i>	49	0	-	-	-	-	-	-
<i>S. fallax</i>	53	0	-	-	-	-	-	-
<i>P. patens</i>	783	0	-	-	-	-	-	-
<i>P. schreberi</i>	9	7	#4 (100)	Синглеты	42.85	98.5	96.7	0
<i>C. plumiforme</i>	24	8	#4 (100)	Синглеты	87.50	93.9	92.8	0
<i>C. purpureus</i>	952	5	#4 (100)	1	25.0	99.7	100.0	1
<i>F. antipyretica</i>	381	0	-	-	-	-	-	-
<i>S. caninervis</i>	119	12	#4 (91.7)	1	33.3	90.7	95.3	0
<i>S. tamariscina</i>	303	73	#1 (20.8)	1	100.0	98.6	99.2	0
			#4 (70.8)	3	60.0	95.8	99.4	0
<i>S. moellendorffii</i>	82	51	#2 (100)	2	40.7	93.1	99.5	0
<i>S. kraussiana</i>	3	2	#2 (50.0)	Синглеты	0	-	-	-
			#3 (50.0)	Синглеты	0	-	-	-
<i>I. engelmannii</i>	116	110	#2 (88.1)	3	35.7	-	96.7	0
<i>Polypodium sp</i>	0	0	-	-	-	-	-	-

<i>Pteridium sp</i>	9	3	#1 (100)	Синглеты	33.33	-	-	-
<i>Plagiogyria sp</i>	8	9	#1 (100)	1	50.0	-	-	-
<i>Dipteris sp</i>	5	0	-	-	-	-	-	-
<i>Cystopteris sp</i>	5	4	#1 (100)	Синглеты	0	-	-	-
<i>Ceratopteris sp</i>	0	0	-	-	-	-	-	-
<i>C. richardii</i>	3283	53	#1 (100)	2	30.0	90.9	91.1	0
<i>S. cucullata</i>	46	17	#1 (100)	1	0.0	91.9	94.3	0
<i>A. filiculoides</i>	2436	167	#1 (100)	2	30.4	98.5	100.0	23
<i>G. montanum</i>	3	26150	#1 (88.7)	58	1.6	88.9	100.0	3
			#2 (11.1)	17	11.6	92.4	98.6	0
<i>W. mirabilis</i>	21	879	#1 (90.0)	15	46.7	88.1	98.2	0
			#2 (0.6)	Синглеты	0	-	-	-
<i>G. biloba</i>	19833	21357	#1 (23.5)	22	54.7	92.0	99.2	0
			#2 (71.5)	26	15.2	88.8	100.0	1
			#3 (3.8)	10	79.8	89.3	97.6	0
<i>C. panzhihuaensis</i>	29603	2906	#2 (10.7)	7	30.6	82.1	87.1	0
			#3 (68.0)	10	72.4	89.8	98.5	0
<i>P. menziesii</i>	37432	65496	#2 (66.2)	61	18.6	93.2	100.0	2
			#3 (31.6)	40	80.4	92.1	100.0	1
<i>A. alba</i>	42997	86276	#2 (73.2)	148	20.4	90.9	100.0	13
			#3 (25.2)	75	73.3	91.4	98.7	0
<i>P. taeda</i>	24906	39538	#2 (78.5)	64	29.9	91.3	100.0	13
			#3 (18.7)	30	72.0	91.7	100.0	2
<i>P. lambertiana</i>	24900	49880	#2 (82.9)	160	20.5	91.6	100.0	3
			#3 (14.9)	34	70.0	91.1	98.2	0
<i>P. abies</i>	11967	29708	#2 (86.3)	37	57.4	90.6	100.0	2
			#3 (12.5)	12	80.5	88.9	98.1	0
<i>L. sibirica</i>	170	2135	#2 (93.2)	13	30.8	88.0	100.0	1
			#3 (1.6)	3	25.0	-	-	-
<i>P. sitchensis</i>	29162	50267	#2 (85.7)	63	41.7	90.1	100.0	23
			#3 (13.5)	14	65.9	89.4	95.1	0

<i>L. kaempferi</i>	31175	27181	#2 (50.2)	50	22.9	92.7	100.0	5
			#3 (47.2)	61	59.3	92.5	100.0	2
<i>P. glauca</i>	6744	21831	#2 (80.3)	36	28.2	91.5	100.0	68
			#3 (17.9)	14	68.1	93.5	100.0	14
<i>S. giganteum</i>	169	56032	#1 (65.0)	109	67.1	90.2	100.0	125
			#2 (34.7)	36	34.5	90.5	100.0	54
<i>S. sempervirens</i>	2409	164834	#1 (59.3)	334	69.5	91.4	100.0	66
			#2 (40.2)	106	38.2	92.8	100.0	25
<i>T. plicata</i>	1769	49262	#1 (63.2)	79	67.4	91.9	100.0	12
			#2 (36.3)	40	31.7	92.1	100.0	13
<i>T. wallichiana</i>	25761	73178	#1 (34.3)	281	54.4	92.0	100.0	26
			#2 (64.7)	294	3.4	91.3	99.7	0
<i>A. trichopoda</i>	0	1702	#1 (99.8)	63	30.5	81.4	91.5	0
<i>N. colorata</i>	0	314	#1 (100)	7	26.9	95.2	100.0	1
<i>N. thermarum</i>	0	160	#1 (100)	5	19.0	94.8	99.0	0
<i>A. fimbriata</i>	0	1322	#1 (99.9)	14	72.1	93.1	100.0	1
<i>L. chinense</i>	0	13029	#1 (99.6)	140	89.8	90.0	100.0	61
<i>C. salicifolius</i>	0	1	#1 (100)	Синглеты	0	-	-	-
<i>C. kanehirae</i>	0	1624	#1 (99.0)	36	62.3	93.1	100.0	7
<i>P. americana</i>	0	2862	#1 (99.5)	39	61.5	94.3	100.0	3
<i>L. cubeba</i>	0	10800	#1 (99.4)	39	29.5	97.8	100.0	153