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

На правах рукописи

Бондар Евгения Ивановна

Аннотация генома и предсказание сайтов начала транскрипции для лиственницы сибирской (*Larix sibirica* Ledeb.)

Специальность 1.5.7 — Генетика

Диссертация на соискание ученой степени кандидата биологических наук

Научный руководитель:
Doctor of Philosophy (PhD),
кандидат биологических наук, доцент
Татаринова Татьяна Валерьевна

Красноярск 2023

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования

У более чем 70% видов всех растений не было секвенировано ни одного участка ДНК, не говоря уже о полном геноме [1]. Из 1310 геномов отдельных видов растений, опубликованных в NCBI по состоянию на сентябрь 2023 г., 1256 относятся к покрытосеменным, 24 — к голосеменным, 20 — к мхам, 4 — к папоротникам и 4 — к плаунам [2]. Таксономическое распределение публично доступных геномов растений довольно смещено в сторону сельскохозяйственных культур [3]. Многие опубликованные геномы, хотя и достаточно полные на уровне последовательности, имеют очень фрагментарные сборки. Обилие псевдогенов, увеличенное число генных семейств [4] и пролиферирующая активность мобильных элементов [5,6] затрудняют корректную сборку и аннотацию многих растительных геномов.

Хвойные — древняя группа голосеменных растений. Более 600 видов этой группы имеют важную роль в экосистемах бореальных лесов [7,8]. К их отличительным особенностям, помимо прочего, относятся крайне большие размеры генома, а также высокое содержание повторяющейся ДНК и мобильных элементов, что делает расшифровку таких геномов более трудоемкой и затратной по времени, чем у других растений. Несколько мегагеномов хвойных видов были недавно секвенированы и собраны до чернового состояния [9–17], что позволяет, несмотря на их неполноту, уже сейчас проводить структурный и функциональный анализ. Неполный геном также может быть ценным источником данных для понимания регуляторных отношений между элементами генома.

Лиственница сибирская (*Larix sibirica* Ledeb.) — листопадное хвойное дерево, является одним из главных компонентов хвойных лесов, и занимают около 40% лесистой территории России [18-19]. Этот вид отличается высокой устойчивостью к низким температурам и гниению древесин, а также быстрым ростом, что делает его особенно ценным для использования в строительстве. Геном лиственницы

сибирской был впервые опубликован в 2019 году [12], а её аннотация получена в представленном здесь исследовании [20].

Степень разработанности темы

Благодаря быстро развивающимся технологиям высокопроизводительного секвенирования были секвенированы и опубликованы геномы для одиннадцати видов хвойных в семействе Pinaceae, включая ель обыкновенную (*Picea abies* (L.) Karst.) [17], ель белую (*P. glauca)* [16], сосну ладанную (*Pinus taeda* L.) [13,21,22], сосну сахарную (*Pinus lambertiana* Douglas) [23], псевдотсугу Мензиса (*Pseudotsuga menziesii* (Mirb.) Franco) [14], пихту белую (*Abies alba* Mill.) [11], лиственницу сибирскую (*Larix sibirica),* лиственницу японскую (*Larix kaempferi* (Lamb.) Carr.) [9], сосну красную китайскую (*Pinus tabuliformis* Carr.) [10], ель Энгельмана (*Picea engelmannii* Parry ex Engelm.) (NCBI BioProject PRJNA504036) и ель ситхинскую (*Picea sitchensis* (Bong.) Carr.) (NCBI BioProject PRJNA304257).

Важное экологическое и экономическое значение лиственницы сибирской стимулировало изучение ее популяционной структуры [24–26] и разработку молекулярно-генетических маркеров [27,28]. Полногеномное секвенирование позволило разработать дополнительные высокоинформативные видоспецифичные микросателлитные маркеры (т.н. SSR-маркеры от «simple sequence repeats») *L. sibirica* [29,30], которые могут использоваться для различных практических целей, включая борьбу с незаконными рубками [31]. Публикация первых ядерного [12], хлоропластного [32] и митохондриального [33] геномов лиственницы сибирской, а недавно и лиственницы Кемпфера [9] значительно способствовали развитию геномного ресурса для рода *Larix*.

Цели и задачи исследования

Основной целью данного исследования было получение аннотации полного генома лиственницы сибирской *Larix sibirica* Ledeb., а также ее улучшение с помощью полногеномного предсказания сайтов начала транскрипции.

Исходя из поставленной цели, были сформулированы следующие задачи:

1. проанализировать относительное содержание высокоповторяющихся элементов в геномной сборке лиственницы сибирской;
2. выполнить структурную и функциональную аннотацию генов для лиственницы сибирской и сравнить с имеющимися аннотациями для других видов семейства Pinaceae;
3. предсказать *de novo* сайты начала транскрипции (transcription start sites, TSS) для генома лиственницы и других видов хвойных;
4. разработать видоспецифичные SSR-маркеры для лиственницы сибирской.

Научная новизна исследования

Впервые представлена подробная аннотация генов и мобильных элементов генома лиственницы сибирской. Данная аннотация является первым публично доступным ресурсом для рода *Larix*. Была получена комплексная видоспецифичная библиотека повторов, которая может использоваться для поиска мобильных элементов в геномах других голосеменных растений. Были разработаны и протестированы полиморфные SSR-маркеры для лиственницы сибирской, подходящие также для популяционно-генетических исследований лиственниц Гмелина и Каяндера. С помощью компьютерных методов, основанных на максимизации ожидания и нейронных сетях, были предсказаны сайты начала транскрипции для трех видов семейства Pinaceae. Для проверки точности предсказаний был использован метод валидации *de novo*, основанный на распределении длин 5′-нетранслируемой области, профиле распределения свободной энергии ДНК дуплексов и позиционном распределении сайтов связывания транскрипционных факторов. Эта работа является первым полногеномным предсказанием TSS в геномах размером более 10 млрд. п.н.

Теоретическая и практическая значимость работы

Теоретическая значимость работы обусловлена тем, что черновые сборки и аннотации геномов хвойных являются ценным ресурсом для дальнейших генетических и геномных исследований. Существующие аннотации геномов хвойных позволяют выявлять различия между голосеменными и

покрытосеменными видами на уровне генома, проявляющиеся в различной представленности генов в функциональных категориях.

Разработанные в данной работе полиморфные SSR маркеры позволяют оценивать уровень генетического разнообразия и дифференциации популяций лиственницы сибирской. Данные маркеры могут успешно применяться для изучения так же лиственниц Гмелина и Каяндера. Они разработаны на основе длинных три-, тетра- и пентануклеотидных мотивов, и поэтому могут анализироваться при помощи простого гель-электрофореза, в том числе в лабораториях, где отсутствует техническая возможность проведения капиллярного электрофореза.

Идентификация TSS и соответствующих промоторных областей является важным ресурсом для экспериментальной проверки и понимания регуляции генов, а также для исследования эволюционных связей между голосеменными и покрытосеменными растениями. Эта информация может быть использована в генетической селекции и редактировании генома для более точного картирования функциональных областей генома и локусов количественных признаков (QTL), таких как скорость роста, устойчивость к холоду и засухе, резистентность к патогенам и инвазии.

Все данные, полученные в данной работе, включая файлы аннотации и комплексную библиотеку повторов, доступны публично в figshare (DOI 10.6084/m9.figshare.19785913) и в репозитории суперкомпьютерного центра СФУ. Геномные последовательности, треки с генными моделями, предсказания TSS и данные о покрытии РНК-секвенирования доступны в геномном браузере Persephone (https://web.persephonesoft.com).

Положения выносимые на защиту

1. Получена подробная структурная и функциональная аннотация ядерного, митохондриального и хлоропластного геномов для вида *Larix sibirica*. Размер митохондриального генома составил 11,7 млн.п.н., что на текущий момент является самым большим митогеномом из известных.

2. Оценка доли повторов в геноме лиственницы составляет порядка 66%. Вероятный период массированного встраивания ретротранспозонов в геном лиственницы может быть оценен порядка 4–5 млн лет назад.

3. 14 полиморфных микросателлитных маркеров, разработанных в данном исследовании для лиственницы сибирской, могут так же использоваться для популяционно-генетических исследований, Гмелина и Каяндера.

Апробация результатов исследования

Данные по теме диссертации докладывались на ежегодных заседаниях кафедры геномики и биоинформатики СФУ в 2018–20122 г. Материалы диссертации представлены в 6 статьях, включая 5 опубликованных в международных рецензируемых изданиях, индексируемых в базах Web of Science и Scopus, а также в 14 тезисах международных и всероссийских конференций. Доклады по теме диссертации проводились на ежегодных заседаниях кафедры геномики и биоинформатики ИФБиБТ СФУ в 2018-2022 г. Промежуточные и итоговые результаты работы были представлены на российских и международных конференциях: VII Международная научная конференция «Генетика, геномика, биоинформатика и биотехнология растений» («PlantGen2023», 10–15 июля 2023 г., Казань), III Всероссийская конференция «Высокопроизводительное секвенирование в геномике» (19–24 июня 2022 г, Новосибирск), 6-ая Международная научная конференция «Plant Genetics, Genomics, Bioinformatics, and Biotechnology («PlantGen2021», 14–18 июня 2021 г, Новосибирск), Международная конференция американской ассоциации RASA Global (2020, online), 12-ая международная конференция «Биоинформатика регуляции и структуры генома\системная биология BGRS» (06-10 июля 2020 г, Новосибирск), 6-я международная конференция-совещание «Сохранение лесных генетических ресурсов» (16-20 сентября 2019 г, Щучинск, Казахстан), 11-я международная конференция «Биоинформатика регуляции и структуры генома\системная биология BGRS» (2018 г, Новосибирск).

Публикации по теме

1. **Bondar, E. I.** Annotation of Siberian Larch (*Larix sibirica* Ledeb.) Nuclear Genome – One of the Most Cold-Resistant Tree Species in the Only Deciduous GENUS in *Pinaceae* / E. I. Bondar, S. I. Feranchuk, K. A. Miroshnikova, V. V. Sharov, D. A. Kuzmin, N. V. Oreshkova, K. V. Krutovsky // Plants. – 2022a. – Vol. 11, Iss. 15. – P. 2062.
2. **Bondar, E. I**. Genome-wide prediction of transcription start sites in conifers / E. I. Bondar, M. E. Troukhan, K. V. Krutovsky, T. V. Tatarinova // International Journal of Molecular Sciences. – 2022b. – Vol. 23, Iss. 3. – P. 1735.
3. Putintseva, Yu. A. Siberian larch (*Larix sibirica* Ledeb.) mitochondrial genome assembled using both short and long nucleotide sequence reads is currently the largest known mitogenome / Y. A. Putintseva, **E. I. Bondar**, E. P. Simonov, V. V. Sharov, N. V. Oreshkova, D. A. Kuzmin, Y. M. Konstantinov, V.N. Shmakov, V.I. Belkov, M.G. Sadovsky, O. Keech, K. V. Krutovsky // BMC genomics. – 2020. – Vol. 21, Iss. 1. – P. 1-12.
4. **Bondar, E. I**. Siberian larch (*Larix sibirica* Ledeb.) chloroplast genome and development of polymorphic chloroplast markers / E. I. Bondar, Y. A. Putintseva, N. V. Oreshkova, K. V. Krutovsky // BMC bioinformatics. – 2019. – Vol. 20, Iss. 1. – P. 47-52.
5. Орешкова, Н. В. Разработка ядерных микросателлитных маркеров с длинными (трех-, четырех-, пяти- и шестинуклеотидными) мотивами для трех видов лиственницы на основе полногеномного *de novo* секвенирования лиственницы сибирской (*Larix sibirica* Ledeb.) / Н. В. Орешкова, **Е. И. Бондар**, Ю. А. Путинцева, В. В. Шаров, Д. А. Кузьмин, К. В. Крутовский // Генетика. – 2019. – Т. 55, № 4. – С. 418-425.

Личный вклад автора в проведенные исследования

Автором работы выполнены лично: аннотация хлоропластного генома, проверка сборки и анализ повторов в митохондриальном геноме, черновая аннотация митохондриального генома, идентификация и анализ повторов в

ядерном геноме, оценка эволюционного времени расхождения (дивергенции) или вставки дуплицированных длинных концевых повторов-ретротранспозонов (LTR-RT) внутри вида, функциональная аннотация ядерного генома, сравнительный анализ представленности категорий генной онтологии, предсказание сайтов начала транскрипции и анализ статистических свойств геномов хвойных, подготовка данных к публикации и написание рукописей статей.

Секвенирование транскриптома лиственницы сибирской проводилось сотрудниками лаборатории лесной геномики СФУ в рамках гранта Правительства РФ (договор № 14.Y26.31.0004). Сборка транскриптомных данных выполнена сотрудником лаборатории лесной геномики СФУ Бирюковым В.В. Секвенирование и сборка митохондриального генома лиственницы сибирской проводились в рамках гранта РФФИ № 16-04-01400, выделение интактных митохондрий и обогащенной митохондриальной ДНК проводилась в СИФИБР СО РАН в лаборатории генетической инженерии растений под руководством Константинова Ю.М., работы по аннотации митохондриального генома выполнены совместно с Путинцевой Ю.А. Идентификация генов с лейцин богатыми повторами проводилась Мирошниковой К.А. Организация запуска пайплайна для аннотации на 448 ядерном вычислительном кластере СФУ проводилась совместно с сотрудниками кафедры высокопроизводительных вычислений СФУ под руководством Кузьмина Д.А. и Феранчука И.С. Образцы лиственницы сибирской для тестирования микросателлитных локусов предоставлены сотрудниками отдела мониторинга состояния лесных генетических ресурсов Центра защиты леса г. Красноярска.

Структура и объём диссертации

Диссертация состоит из введения, обзора литературы, материалов и методов, результатов и их обсуждения, заключения, выводов, списка сокращений и условных обозначений; списка литературы (360 источников) и 7 приложений. Общий объём составляет 151 страницу, содержит 26 рисунков и 13 таблиц.

Благодарности

Автор выражает глубокую благодарность научному руководителю PhD Татариновой Т.В. и руководителю лаборатории лесной геномики к.б.н. Крутовскому К.В. за неоценимую помощь на всех этапах работы.

Отдельную благодарность автор выражает заведующей лаборатории геномных исследований и биотехнологии ФИЦ КНЦ СО РАН к.б.н. Орешковой Н.В. за помощь в тестировании микросателлитных маркеров и обработке результатов генотипирования, заведующему кафедрой высокопроизводительных вычислений к.т.н. Кузьмину Д.А., научному сотруднику ФИЦ КНЦ СО РАН Шарову В.В, и к.ф.-м.н. Феранчуку И.С. за помощь в вычислениях и обработке данных; Путинцевой Ю.А. за помощь в сборке и аннотации хлоропластного генома. Автор благодарен сотрудникам отдела мониторинга состояния лесных генетических ресурсов Центра защиты леса г. Красноярска за предоставленные образцы лиственницы сибирской.

Автор выражает признательность заведующей кафедрой геномики и биоинформатики к.б.н. Ямских И.Е. и ведущему научному сотруднику ИВМ СО РАН д.ф.-м.н. Садовскому М.Г. за ценные комментарии, а также старшему научному сотруднику к.б.н. Клепиковой А.В., научному сотруднику ИЦиГ СО РАН к.б.н. Дорошкову А.В. и заведующему лабораторией популяционной генетики ИОГен РАН д.б.н. Политову Д.В. за рецензирование работы.

Также автор признателен Мирошниковой К.А., Бирюкову В.В., Акуловой В.С., Новиковой С.В. и Тараненко Е.А. за понимание и поддержку в время работы над диссертацией.

Диссертационная работа выполнена на базе кафедры геномики и биоинформатики и лаборатории лесной геномики СФУ в рамках проекта «Геномные исследования основных бореальных лесообразующих хвойных видов и их наиболее опасных патогенов в Российской Федерации», финансируемого Правительством РФ (договор №14.Y26.31.0004, руководитель проекта проф. К. В. Крутовский), а так же в рамках гранта РФФИ № 16-04-01400 под руководством проф. К. В. Крутовского.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

Аннотирование генома — это процесс идентификации функциональных элементов в последовательности ДНК. Аннотация придает смысл геному, указывая расположение и функцию генов (кодирующих белок или выполняющих иные функции), а также регуляторных элементов и областей, лежащих в основе биологии живых организмов. Наличие известного набора функциональных областей генома облегчает последующий анализ, такой как поиск дифференциально экспрессирующихся генов, определение локусов количественных признаков, поиск ассоциации между однонуклеотидными полиморфизмами и изменчивостью отдельных признаков [34].

1.1 Обзор публично доступных геномов и аннотаций высших растений

Представители клады высших (эукариотических или зеленых) растений (Viridiplantae) являются фундаментальной движущей силой глобальных экосистем, внося огромный вклад в сельское хозяйство, медицину и естественные экологические процессы [35]. На сегодняшний день численность видов в этой группе составляет почти полмиллиона [1,3]. Поскольку проблемы разрушения среды обитания, потери видов, изменения климата и резкого изменения взаимодействия сообществ стоят как никогда остро, информация о структуре и свойствах геномов представителей растительного мира имеет крайне важное значение для поиска решений тех фундаментальных проблем, с которыми сегодня сталкивается человечество. Несмотря на это, наши текущие знания об организации, структуре и функциональных особенностях растительных геномов значительно отстают от таковых в области, например, геномики позвоночных, для которых сейчас разрабатывается и тестируется основная масса молекулярно-генетических протоколов и биоинформатических алгоритмов [1].

Геномы многих представителей царства животных обычно относительно компактны, но их размеры широко варьируют от 19,6 млн. п.н. у нематоды (*Pratylenchus coffeae* Goodey) до 132,83 млрд. п.н. у двоякодышащей рыбы

мраморного протоптера (*Protopterus aethiopicus* Heckel) [36,37]. Схожим образом, для наземных растений размеры генома варьируют от ~60–64 млн. п.н. для некоторых видов рода *Genlisea* [38] до 148,9 млрд. п.н. у париса японского (*Paris japonica* (Thunb.) D.Don ex G.Don) [39], обладателя самого большого из известных эукариотических геномов. За такую изменчивость размера генома у растений чаще всего ответственны два механизма: неоднократные эпизоды полногеномной дупликации (whole genome duplication — WGD или полиплоидия), распространенные у растений, и динамика встраивания-элиминации мобильных генетических элементов — транспозонов и ретротранспозонов (Рисунок 1Б). Количество белок-кодирующих генов в растительных геномах относительно стабильно и по некоторым оценкам составляет около 40 000 [40], однако также может варьировать от 27 655 у модельного *Arabidopsis thaliana* (L.) Heynh. [41], до 50 894 у тетраплоидной люцерны (*Medicago trunculata* Gaertn.) [42], 74 350 у тетраплоидного хлопчатника (*Gossypium hirsutum* L.) [43] и 80 495 у китайской красной сосны (*P. tabuliformis*) [10] (Рисунок 1А). Также для черновых сборок нескольких хвойных видов число генов оценено в 94 205 для *A. alba* [11], 102 915 для *P. glauca* [16] и 118 906 для *Sequoia sempervirens* (D.Don) Endl. [44].

Для более чем 70% всех видов растений не было секвенировано ни одного участка ДНК, не говоря уже о полном геноме [1]. Из 1310 геномов отдельных видов растений, опубликованных в NCBI по состоянию на сентябрь 2023 г., 1256 относятся к покрытосеменным, 24 — к голосеменным (11 Pinacea, 7 Cupressaceae, *2* Taxaceae, Cycas, Ginkgo, Gnetum, Welwitschia), 20 — к мхам, 4 — к папоротникам и 4 — к плаунам [2]. Таксономическое распределение доступных геномов довольно асимметрично и, по понятным причинам, смещено в сторону сельскохозяйственных культур. Так, например, самыми хорошо представленными семействами являются Poaceae (114 видов), Brassicaceae (105 вид), Fabaceae (87 вид) и Solanaceae (77 видов). Наличие значимой информации о геноме для 1416 видов растений — это лишь малая часть из более чем 412 000 существующих видов высших (зеленых) растений [3]. Таксономический архив NCBI содержит порядка 247 376 видов высших (зеленых) растений (по состоянию на сентябрь 2023 г.) [45],

что составляет лишь около половины предполагаемого видового разнообразия высших растений. Для многих из этих видов было секвенировано лишь ограниченное количество генов. Более того, истинная численность видов внутри некоторых из этих таксономических групп по большей части остается неизвестной [1].

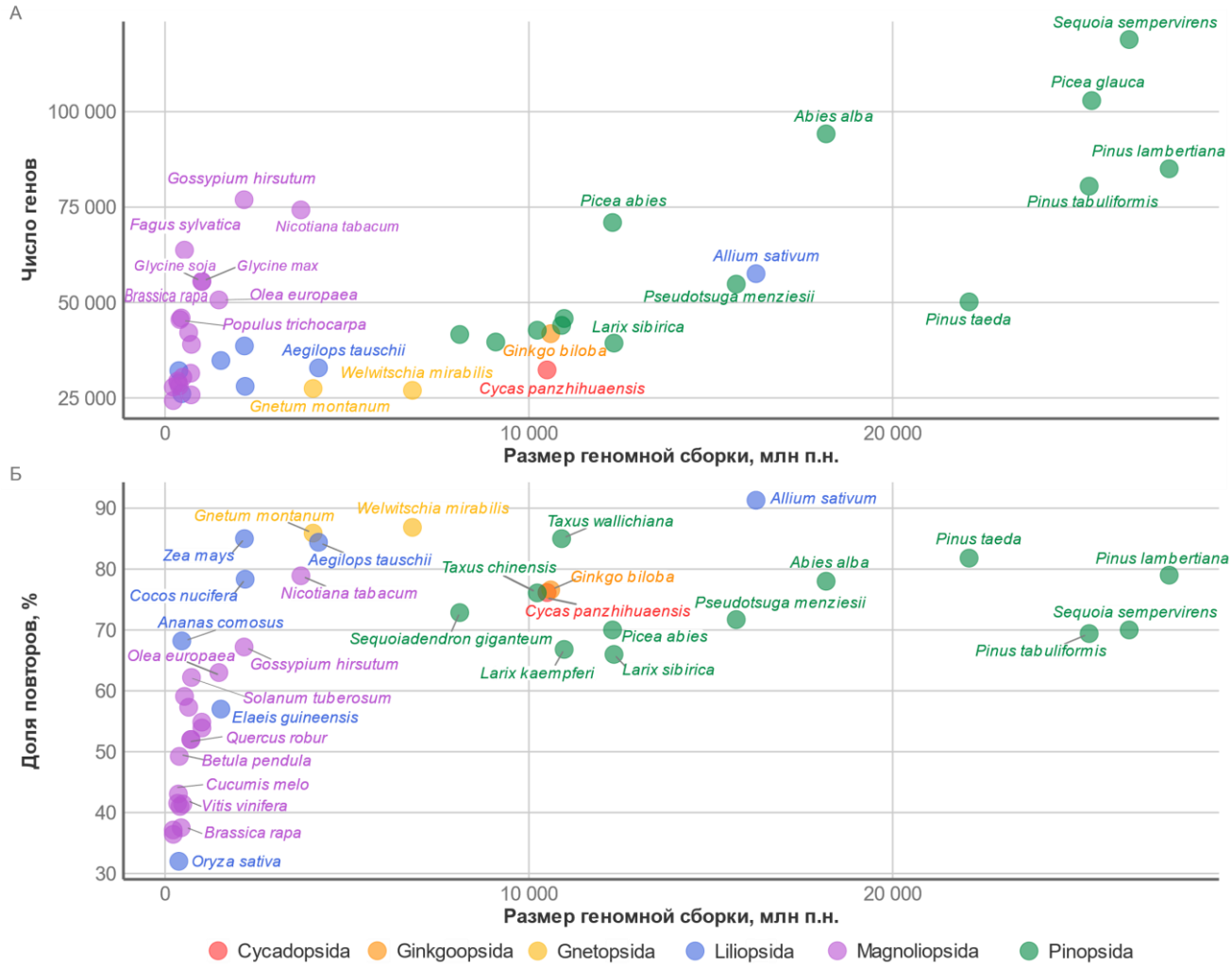


Рисунок 1. Геномы семенных растений, секвенированных и аннотированных к 2023 году: **А** — отношение размеров генома к общему числу генов; **Б** — отношение размеров генома к доле повторов

Голосеменные возникли примерно 360 млн. лет назад, когда они составляли преобладающую часть наземной растительности на Земле [7,46]. Ныне живущие голосеменные растения насчитывают около 1 000 видов [7], причем хвойные представляют собой самую разнообразную и многочисленную группу. Являясь одними из древнейших семенных растений, они рассматриваются как связующее звено между покрытосеменными и папоротниковыми. Благодаря быстро

развивающимся высокопроизводительным технологиям были секвенированы геномы для 24 видов голосеменных растений, включая 11 видов в семействе Pinaceae, 7 видов в семействе Cupressaceae, 2 вида в семействе Taxaceae, а также геномы *Cycas panzhihuaensis* L. Zhou & S. Y. Yang, *Gnetum montanum* Markgr., *Welwitschia mirabilis* Hook.f. и *Ginkgo biloba* L. (Таблица 1).

Таблица 1. Статистика секвенированных геномов голосеменных растений

Семейство	**Вид**	**Размер сборки, млрд. п.н.**	**Число генов**	**Повторы, %**	**Источник**
Welwitschiaceae	*Welwitschia mirabilis*	6,8	26 990	87	Wan et al. (2021)
Gnetaceae	*Gnetum montanum*	4,1	27 491	86	Wan et al. (2018)
Cycadaceae	*Cycas panzhihuaensis*	10,5	32 353	76	Liu et al. (2022)
Ginkgoaceae	*Ginkgo biloba*	10,6	41 840	77	Guan et al. (2016)
Cupressaceae	*Thuja plicata*	9,1	39 659	–	DOE-JGI, Project ID: 1191975
	Sequoiadendron giganteum	8,1	41 632	73	Scott et al. (2020)
	Sequoia sempervirens	26,5	118 906	70	Neale et al. (2022)
	Chamaecyparis obtusa	8,5	5 814	–	Shirasawaa et al. (2023)
	Cryptomeria japonica	9,2	34 468	–	
	Cunninghamia lanceolata	11,7	53 750	–	
	Cupressus sempervirens	10,0	–	–	NCBI PRJNA833966, WGS project JAPHAW01
Taxaceae	*Taxus chinensis*	10,2	42 746	76	Xiong et al. (2021)
	Taxus wallichiana	10,9	44 008	85	Cheng et al. (2021)
Pinaceae	*Abies alba*	18,2	94 205	78	Mosca et al. (2019)
	Larix kaempferi	10,9	45 828	67	Sun et al. (2022)
	Larix sibirica	12,3	39 370	66	Kuzmin et al. (2019)
	Picea abies	12,3	70 968	70	Nystedt et al. (2013)
	Picea glauca	25,3	102 915	–	Birol et al. (2013); Warren et al. (2015)
	Picea engelmannii	24,9	–	–	NCBI PRJNA504036
	Picea sitchensis	20,5	–	–	NCBI PRJNA304257
	Pinus lambertiana	27,6	85 053	79	Gonzalez-Ibeas et al. (2016)
	Pinus tabuliformis	25,4	80 495	69	Niu S. et al. (2022)
	Pinus taeda	22,1	50 172	82	Neale et al. (2014); Zimin et al. (2017)
	Pseudotsuga menziesii	15,7	54 830	72	Neale et al. (2017)

Геномы хвойных голосеменных имеют ряд особенностей, отличающих их от других растений, самая примечательная из них — огромный размер, не являющийся результатом полиплоидизации, по крайней мере, недавней. Он варьирует среди секвенированных видов от 4 млрд. п.н. у *G. montanum* [47] до 31 млрд. п.н. у сахарной сосны *P. lambertiana* [15,23], что намного больше, по

сравнению с типичными диплоидными покрытосеменными растениями, например, 135 млн. п.н. у двудольного *A. thaliana* [48] или 3,1 млрд. п.н. у двудольного подсолнечника (*Helianthus annuus* L.) [49], но сопоставимо с некоторыми полиплоидными покрытосеменными, например, 14,5 млрд. п.н. у мягкой пшеницы (*Triticum aestivum* L.) [50] или 150 млрд. п.н. у Париса японского (*P. japonica)* [39]. У секвенированных хвойных также, вероятно, различается и количество генов, так как количество предсказанных генных моделей варьирует от 26 990 у *W. mirabilis* [51] и 32 353 у саговника паньчжихуанского (*C. panzhihuaensis)* [52] до 102 915 у ели белой (*P. glauca)* [16] и 118 906 у секвойи вечнозелёной (*S. sempervirens)* [44].

Было показано, что разница в размере генома хвойных не связана с недавней полиплоидизацией или дупликацией всего генома [53], однако число копий генов у голосеменных выше, чем у большинства покрытосеменных видов, что может быть связано с транспонированной дупликацией («копирование» гена из исходного местоположения в новое, «транспонированное») и диспергированной дупликацией генов (создание двух копий генов, которые не являются ни соседними, ни коллинеарными; происходит по всей видимости случайным образом, механизм их появления неясен) [4]. Еще одной характеристикой геномов хвойных является большая доля повторяющейся ДНК, составляющая, по разным оценкам, 70–82% от размера генома [14,21,54,55]. Предполагается, что именно вставки и экстенсивная пролиферация мобильных элементов были в основном ответственны за увеличение размеров генома хвойных [5].

Многие опубликованные геномы растений, хотя достаточно полные на уровне последовательности, имеют очень фрагментарные сборки. Из-за обилия псевдогенов, увеличения числа генных семейств и пролиферирующей активности мобильных элементов корректная сборка и аннотация многих растительных геномов сильно осложнена [6]. Учитывая то, что связь между размером генома растения и содержанием мобильных элементов обычно является линейной в пределах одного уровня плоидности [56], вариации в динамике процессов пролиферации-элиминации мобильных элементов являются самым большим источником вариаций размера генома растений, что делает их намного более

сложными, чем геномы позвоночных [1]. Стоит отметить, однако, что качество и полнота сборок начинают значительно улучшаться в последние годы за счёт большего использования длинных прочтений в сочетании с новыми методами 3D геномики.

1.2 Краткая характеристика вида лиственницы сибирской

Лиственница сибирская была впервые описана как самостоятельный вид в 1832 г. Карлом Фридрихом Ледебуром (Carl Friedrich von Ledebour) в его «Flora altaica» в ходе первой ботанической экспедиции, обследовавшей весь Алтай. До этого Петром Палласом она относилась к *Pinus larix* Pall., т.е. к тому же виду, к которому принадлежала лиственница, растущая в Европе. В настоящее время большинство исследователей под именем *Larix sibirica* понимают лиственницу, произрастающую в северо-восточной части Европейской России, на Урале, в Западной и части Восточной Сибири, на Алтае, в Саянах и Монголии [57].

Лиственница сибирская это высокое, до 35–40 м, дерево, с глубоко идущим стержневым корнем, что делает ее ветроустойчивой, с хорошо развитыми боковыми корнями. Крона у лиственницы сибирской обычно яйцевидно-конической формы. Кора у взрослых деревьев толстая, глубоко-бороздчатая, серовато-бурая, отделяющаяся крупными кусками. Хвоя мягкая, узколинейная, ярко-зелёная с сизоватым налётом, одиночная на удлиненных побегах и сидящая пучками на укороченных, прямая или слегка серповидно-изогнутая. На укороченных побегах располагаются мужские микростробилы и женские шишки. Женские шишки 2,5–4 см длиной, яйцевидные, с ложкообразными опушёнными семенными чешуями (**Рисунок 2**). Опыление у *L. sibirica* начинается одновременно с развертыванием листьев в конце апреля и продолжается до середины мая. Семена у лиственницы сибирской могут достигать 6 мм в длину, крылатые (крыло 8–17 мм), с твердой оболочкой, желто-коричневого цвета. В популяциях, произрастающих южнее, время рассеивания семян приходится на осень; в растущих в средней и северной части ареала — на февраль–март. При высокой влажности шишки раскрываются позже, вплоть до лета следующего года. После

выпадения семян шишки могут оставаться на дереве до трёх лет. Живет лиственница около 300–450 лет, предельный возраст составляет 900 лет [57].

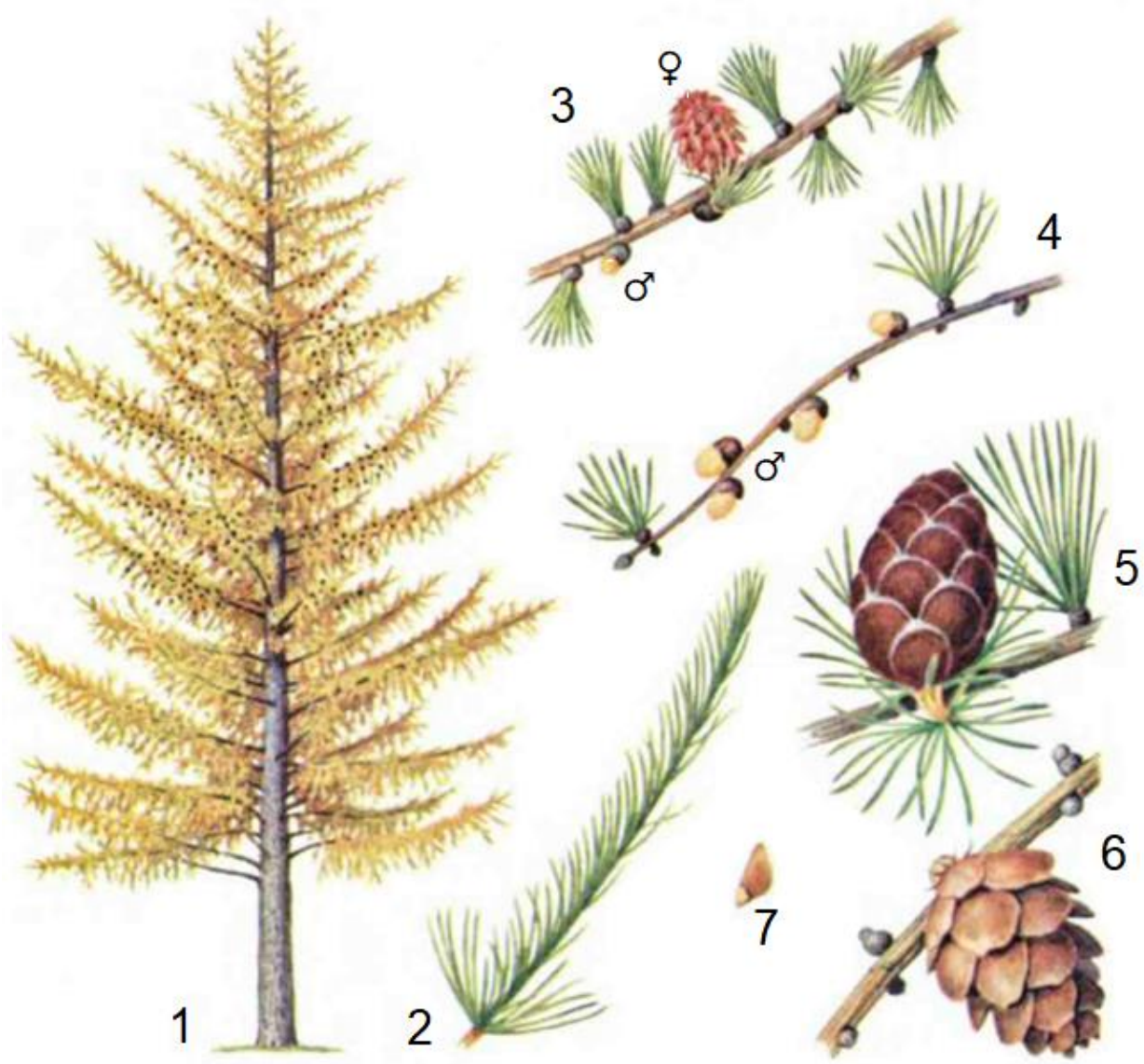


Рисунок 2. *Larix sibirica* [58]. 1 — общий вид дерева (осеняя окраска); 2 — ростовой побег; 3 — ветвь с макро- и микростробилами; 4 — ветвь с микростробилами; 5 — сформировавшаяся шишка; 6 — зрелая шишка; 7 — семя с крылом.

На протяжении своего обширного ареала лиственница сибирская далеко неоднородна, сильно варьирует по целому ряду признаков. Внутри вида нередко выделяют по крайней мере несколько географических рас или экотипов. Ареал произрастания расположен в западной и восточной частях севера России, а также в Сибири. На юге она встречается на территории Урала, Алтая и Саян, а на западе граничит с ареалом лиственницы Сукачева [59]. В горных районах высотность мест ее произрастания может доходить до 2100–2500 м над уровнем моря [19]. Лиственница сибирская является господствующим видом в центральной и юго-

западной частях Алтая, в его засушливых горных степных и полупустынных районах порой является единственной древесной породой [60].

Лиственница сибирская имеет специфические генетические механизмы адаптации, позволяющие ей выживать в суровых условиях холодного севера, в том числе способность сбрасывать хвою в зимний период («листопадность»), что среди семейства сосновых Pinaceae характерно только для лиственниц [61]. При достаточном освещении ее фотосинтетический аппарат обеспечивает высокую интенсивность фотосинтеза и эффективную терморегуляцию за счет энергичной транспирации и интенсивного развития поверхностной корневой системы и придаточных корней — приспособления к произрастанию на мерзлых почвах в условиях короткого периода вегетации [60]. Этот вид является одним из наименее требовательных к климатическим условиям. В пределах ареала лиственница встречается как на многолетней мерзлоте, так и на болотистой местности. Лиственница сибирская может переносить низкие температуры, мириться с коротким летним периодом и со значительной сухостью почвы, однако избегает бедных песчаных почв, предпочитая известняки и гипсы [19].

На долю лиственничных лесов в России приходится наибольшая часть запасов древесины. По техническим характеристикам, скорости роста и качеству древесина лиственницы превосходит древесину всех произрастающих на территории России хвойных видов. Сучковатость лиственницы в два-три раза меньше, чем у других хвойных и твердолиственных пород [62]. Лиственничная древесина очень твердая, плотная и тяжелая, что создает определенные проблемы при ее обработке. Однако благодаря большому количеству смолы она хорошо противостоит гниению и отлично сохраняется в воде, ввиду чего используется в судостроении и на гидротехнических сооружениях, при изготовлении шпал, столбов, свай, водоспусков, а также для отделки. В перерабатывающей промышленности из нее получают целлюлозу, этиловый спирт, камедь [57].

Определенную ценность представляет кора лиственницы, на которую приходится порядка 30% от объема заготовленной древесины. В лиственничной коре содержится до 20% целлюлозы, 47% лигнина, около 7% пентозанов [62]. Кора

лиственницы содержит до 13% таннидов (производных многоатомных фенолов); экстракт коры используется в качестве дубителя и красителя, окрашивающего в желтые, розоватые и коричневые тона. Из смолы добывают терпентин (скипидар). Хвоя, корни и ветви лиственницы считаются ценным сырьем для выработки эфирных масел — в ветвях содержится 0,2–0,32%, а в хвое — 0,40–0,35% эфирных масел от веса сухого вещества.

Промышленные выбросы негативно сказываются на хвое лиственницы, однако листопадность делает ее более дымо- и газоустойчивой породой, чем другие хвойные — благодаря этому лиственница становится полезной при создании лесопарковых ландшафтов для городского озеленения [19]. Кроме этого, лиственничные леса применяются в полезащитном лесоразведении [62].

К настоящему времени для популяционно-генетических исследований лиственницы сибирской разработан набор из 11 высокополиморфных видоспецифичных динуклеотидных микросателлитных локусов. [63]. Однако применение таких SSR-маркеров требует использования капиллярного электрофореза, которым располагают не все генетические лаборатории. На обычных гель-электрофорезах крайне трудно работать со столь короткими мотивами при различии фрагментов по длине всего на пару нуклеотидов. В то же время, многие подразделения, входящие в состав профильных институтов, подотчетных Федеральному агентству лесного хозяйства, а также лаборатории филиалов Центра защиты леса РФ и лаборатории институтов РАН, специализирующиеся на изучении разнообразия лесов, нуждаются в современных молекулярно-генетических методиках оценки видового разнообразия и мониторинга происхождения древесины. Таким образом, существует объективная потребность в разработке генетических микросателлитных маркеров с более длинными мотивами, с которыми удобно работать в условиях обычной генетической лаборатории, с использованием простого гель-электрофореза.

Полные геномы для рода лиственницы на сегодняшний день доступны только для двух видов, *Larix kaempferi* и *Larix sibirica* [9,12,32]. Отсутствие полной геномной аннотации, а также разнообразного арсенала геномных данных,

становится препятствием для проведения более сложных ассоциативных исследований и разработки маркеров для геномной селекции.

1.3 Общий подход к геномной аннотации

Аннотирование генома остается трудоемким и ресурсоемким процессом, который сочетает в себе многочисленные типы анализа последовательностей и эвристического прогнозирования [64]. Учитывая размер и сложность многих растительных геномов, первым шагом к полногеномной сборке, как правило, является получение данных секвенирования с короткими прочтениями для получения черновой сборки и предсказания генных моделей. Фактически для получения последовательности генома можно полагаться на NGS секвенирование (Next Generation Sequencing, секвенирование следующего поколения), для которого разработаны специализированные методы сборки последовательностей. Этот подход позволяет получить черновую сборку, достаточную для оценки числа генов и содержания повторов, но имеет ограниченную полезность для исследований хромосомной организации и структурных перестроек в геноме. Однако, несмотря на их недостаточность для оценки структурной организации генома, такие черновые сборки способствуют получению первоначальных данных, выявляя сложности, возникающие в результате повышенной копийности генных семейств или полногеномной дупликации. Кроме того, черновые сборки информативны в исследовании эволюции видов — служа источником маркеров для изучения филогении, они помогают генерировать гипотезы для дальнейшего исследования видов и служат важным первым шагом к более цельной последовательности генома.

Одним из ключевых этапов в предсказании генных моделей является не только наличие готовой геномной сборки, но и получение данных РНК-секвенирования из разнообразного набора тканей. Эти данные используются в качестве поддерживающего «доказательства» при комбинировании результатов *ab initio* (или *de novo*) предсказания генов и, как правило, должны быть получены от того же самого вида. Предсказание генов *ab initio* представляет собой идентификацию

генов на основе базовых характеристик последовательности, которые определяют их как генные структуры.

Также немаловажным шагом является идентификация мобильных элементов в геноме. Активные транспозоны и ретротранспозоны в своей структуре содержат открытые рамки считывания и участки, кодирующие ферменты, опосредующие их размножение и встраивание в другие места генома, например, гены капсида, интегразы, обратной транскриптазы и рибонуклеазы у ретротранспозонов класса I с длинными концевыми повторами. В неактивных мобильных элементах такие гены накапливают замены, вставки и делеции. Тем не менее в процессе предсказания генных моделей эти структуры могут быть ошибочно приняты за реальные белок-кодирующие гены хозяина. Поэтому важно выполнять предварительный поиск повторов и их маскировку в геноме. На практике маскировка повторов заключается в записи нуклеотидов в соответствующих участков генома в нижнем регистре (так называемая «мягкая» маскировка) или замене их на символы «N» («жесткая» маскировка).

Весь процесс аннотации состоит, как правило, из следующих этапов:

1) маскирование высокоповторяющихся элементов в геномной последовательности для облегчения аннотации;

2) использование транскриптов и белков одного и того же или родственного вида для верификации *ab initio* предсказаний генных структур;

3) использование алгоритмов поиска генов *ab initio* для идентификации возможных генных структур;

4) комбинирование этих данных для создания первичного набора генных моделей;

5) фильтрация результатов по качеству, для определения наиболее вероятных генных моделей, которые представляют полноразмерные или почти полноразмерные белок-кодирующие участки [34].

К наиболее общим ошибкам относится ложная классификация локусов как ретроэлеменов или псевдогенов, появление пересекающихся координат генных моделей, наличие генов, фрагментированных между несколькими скаффолдами,

пропущенные генные модели и ошибки в определении интрон-экзонной структуры [34]. Длинные интроны (>100 т.п.н.), нередко встречающиеся у видов с большим геномом, могут представлять проблему при конструировании генных моделей, так как не все генные предикторы учитывают возможность наличия таких больших интронов; кроме того, длинные интроны в сочетании с фрагментированной сборкой повышают вероятность «разрыва» генной модели между скаффолдами. Определение точной интрон-экзонной структуры представляет проблему не только для больших и сложных геномов растений. По результатам сравнения всех белок-кодирующих транскриптов и транскриптов длинной некодирующей РНК (lncRNA) в базах RefSeq и Gencode только 27,5% транскриптов Gencode имели точно такие же интроны, как и соответствующие им гены RefSeq [64,65]. Таким образом, даже после 20 лет усилий точная интрон-экзонная структура многих белок-кодирующих генов человека не установлена. Аннотации же большинства других эукариот, за исключением возможно небольших, широко изучаемых модельных организмов, таких как дрожжи или арабидопсис, находятся в еще более незавершенном состоянии [64].

1.3.1 Поиск и маскировка высокоповторяющихся элементов генома

В течение миллиардов лет мобильные элементы встраивались в геномы эукариот, оставляя в них свои копий. Эти копии, как правило, вредны или бесполезны для своих хозяев, следовательно имеют тенденцию эпигенетически подавляться (например, через метилирование) и со временем удаляться из генома. В зависимости от относительной скорости встраивания и удаления мобильных элементов из генома посредством случайных делеций, часть генома, состоящая из встроившихся в разное время повторов, может варьфировать у разных видов, занимая до 84% у некоторых злаков и до 81% у сосны ладанной *P. taeda* [13,55]. Доля повторов растет линейно для геномов размером до 10 млрд. п.н., а затем выходит на плато [66] (**Рисунок 1Б**).

На первый взгляд, идентификация повторов и последующее построение консенсусной последовательности, представляющей исходный мобильный элемент, кажутся достаточно тривиальной задачей. Если большинство копий

повтора деградируют естественным образом, накопление замен и вставок должно быть случайным, и, при наличии некоторых представлений о паттернах накопления нейтральных мутаций и достаточном количестве копий, поиск исходного мобильного элемента должен сводиться к реконструкции филогенетических отношений между мутирующими копиями одной последовательности. Однако здесь существует много сложностей.

Если копии мобильного элемента накопили 20% замен с момента своей вставки в геном хозяина, расстояние между любыми двумя копиями составляет в среднем 40%. Обнаружение схожих последовательностей с таким низким уровнем сходства требует очень чувствительного выравнивания при сравнении последовательностей друг против друга (self-comparison), что делает процесс непрактично медленным. Чтобы обеспечить более точное выравнивание, программы могут работать с меньшим объемом данных, беря на вход лишь порцию генома, но в таком случае повторы с меньшей копийностью могут остаться незамеченными [67].

Одной из серьезных проблем при нахождении концевых участков повтора является обширная фрагментация. Элементы, встроившиеся относительно давно, как правило, сильно фрагментированы либо из-за частичного удаления фрагментов генома, либо из-за вставок других повторов. У многих видов копии мобильных элементов в основном накапливаются в определенных гетерохроматиновых, бедных генами или межгенных областях генома, отчасти из-за того, что в этих областях их влияние на геном хозяина, скорее всего, является нейтральным. В таких участках общая плотность повторов может приближаться к 100%, из-за вложенных друг в друга вставок повторов [68,69]. Соотношение коротких одиночных элементов и полных копий может быть очень высоким, что затрудняет реконструкцию длинного повтора. Часто об активности ДНК-транспозона свидетельствует только наличие небольших элементов с концевыми инвертированными повторами (TIR) [67]. Рекомбинация между активными мобильными элементами также является обычным явлением. Это особенно верно для длинных концевых ретротранспозонов (LTR), которые пролиферируют через

РНК-интермедиаты. Благодаря этому механизму возникают как химерные LTR, так и идентичные LTR, фланкирующие совершенно разные внутренние последовательности [70].

До недавнего времени считалось, что известные эукариотические мобильные элементы имеют размер от 80 п.н. до 15 тыс. п.н. На сегодняшний день в немодельных организмах обнаружены повторы громадных размеров, чаще всего это LTR-элементы размером до 30 тыс. п.н. у растений [21,55], а также новые повторы размером до 180 тыс. п.н. у рыб [71]. Экстремально длинные повторы представляют проблему для программ, специфичных к идентификации определенных классов повторов, которые чаще всего содержат ограничения на размеры при поиске повторов *de novo* [67].

Классические методы обнаружения повторов разделяются на три группы:

1) методы *de novo/ab initio*, обнаруживающие семейства повторов на основе общего принципа повторяемости подпоследовательности;
2) методы на основе сигнатур, использующие предварительно известные характеристики классов повторов для их обнаружения;
3) методы на основе готовых библиотек, полностью зависящие от предварительно подготовленного набора, содержащего консенсусные последовательности семейств повторов.

Методы *de novo* и основанные на сигнатурах, обычно используются в качестве входных данных для библиотечных методов [67]. Методы *de novo* имеют преимущество перед методами, основанными на сигнатурах, так как могут идентифицировать семейства, не принадлежащие к известным классам повторов или не имеющие одной или нескольких общих известных характеристик, используемых методами на основе сигнатур. Они работают, обнаруживая точные или близко совпадающие повторения последовательности, расширяя эти совпадения и в некоторых случаях группируя их в семейства связанных последовательностей. На анализ всего генома, как правило, способны методы, использующие сравнение последовательностей против самих себя, за которым следует кластеризация для разрешения семейств повторов из данных попарного

выравнивания. Такие подходы требуют значительного времени и вычислительных ресурсов. Точная кластеризация выравниваний при этом бывает затруднена из-за высокой фрагментации и мозаичности, присутствующих в семействах повторов. К программам, работающим на основе выравнивания последовательности самой на себя, относят RECON [72], RepeatFInder [73], RepeatScout [74] и TE_finder [75].

Методы, основанные на сигнатурах, определяют повторы распознавая особенности конкретных классов повторов (концевые инвертированные повторы, прямые повторы, белковые мотивы и т. д.), а также отличительные признаки вставок мобильный элементов, например, наличие дупликации целевого сайта (target site duplication, TSD — короткие геномные последовательности, дуплицированные во время вставки) [67]. Часто необходимо использовать несколько характеристик совместно, чтобы компенсировать низкую специфичность каждой из них. Однако, несмотря на это, методы на основе сигнатур обычно страдают от высокого уровня ложноположительных результатов. LTR-элементы и неавтономные ДНК-транспозоны являются особенно подходящей целью для этой группы методов из-за наличия длинных прямых и инвертированных повторов, фланкирующих интактные копии. Алгоритмы поиска LTR и MITE элементов используют подход на основе сигнатур для идентификации потенциальных совпадений-кандидатов, которые затем дополнительно оцениваются на наличие дупликаций целевого сайта TSD, неповторяющихся фланкирующих последовательностей и мотивов специфичных белковых доменов. К программам, работающим на основе такого подхода, можно отнести LTR_Finder [76], LTRdigest [77], LTRharvest [78], MITE-Hunter [79].

1.3.2 *Ab initio* предсказание генов

Классическая структура прокариотических и эукариотических генов включает стартовые (AUG) и стоп-кодоны (UAA/UAG/UGA), которые сигнализируют об этапах инициации и терминации трансляции белка. Эти кодоны вместе с некоторым нуклеотидным контекстом вокруг них определяют кодирующую последовательность (CDS) экзона или экзонов в гене. В геномах эукариот кодирующая часть гена не обязательно является непрерывной, но напротив, часто

прерывается интронами длиной в несколько сотен или тысяч нуклеотидов [80], которые затем вырезаются в ходе сплайсинга при процессинге мРНК.

Границы интронов в большинстве случаев определяются каноническими (GT-AG, GT на 5'-конце и AG на 3'-конце интрона) и неканоническими (GC-AG и AT-AC) сайтами сплайсинга. Было показано, что в геномах растений частота встречаемости канонического сайта сплайсинга GT-AG составляет 98,7%; доля неканонических 1,2% для GC-AG, 0,06% для AT-AC и 0,09% для других комбинаций нуклеотидов [81,82]. Вырожденность генетического кода приводит к тому, что одной аминокислоте может соответствовать несколько синонимичных кодонов. В кодирующих участках частота встречаемости таких синонимичных кодонов определяет предпочтение кодонов (codon usage bias). Предпочтение кодонов варьирует между видами из-за разницы в силе отбора и скорости накопления мутаций [83], позволяя идентифицировать кодирующие части ДНК в геноме конкретного вида. Эти сигналы — наличие старт- и стоп-кодонов, нуклеотидный контекст, предпочтение кодонов, наличие сайтов сплайсинга — позволяют разрабатывать методы *in silico* предсказания местонахождения и определения интрон-экзонной структуры генов [34].

Так, например первые алгоритмы для *ab initio* предсказания генов, такие как GeneMark [84] и GLIMMER [85] использовали скрытые марковские цепи — стохастическую модель, описывающую последовательность состояний, в которой вероятность следующего состояния зависит только от предыдущего — для фиксации зависимостей между последовательными нуклеотидами и позициями кодонов, определяя таким образом начало и конец гена в нуклеотидном контексте [34]. Для предсказания генов лучше всего подходят марковские цепи пятого порядка, в которых вероятность следующего нуклеотида зависит от пяти предыдущих, так как они фиксируют зависимости между последовательными аминокислотами в белках [34]. После первичной оценки предположительных генных моделей используются алгоритмы, определяющие оптимальную структуру гена.

Идентификация эукариотических генов по сравнению с их прокариотическими аналогами требует больше усилий из-за их длины и прерывистой, интрон-экзонной структуры. Точность же таких предсказаний ниже, чем в прокариотических геномах [34]. При предсказании генов необходимо соблюдать компромисс между чувствительностью (доля нуклеотидов, определенных как кодирующие, среди всех кодирующих), и специфичностью (доля действительно кодирующих нуклеотидов, среди всех предсказанных). Поскольку эти два показателя взаимосвязаны, увеличение одного часто уменьшает другое. Отсутствие правильного начала и конца экзона снижает чувствительность и специфичность предсказания гена на уровне экзона. Точность же для всего гена зачастую бывает еще ниже из-за сложности определения правильной комбинации экзонов в геноме. При этом моноэкзонные и короткие гены, кодирующие малые белки [86], часто выбрасываются из конечной аннотации, поскольку в большинстве алгоритмов применяется порог минимальной длины для уменьшения доли ложноположительных предсказаний. Обычные методы предсказания белок-кодирующих генов испытывают сложности с предсказанием и верификацией генов малых белков, и мало эффективных алгоритмов было разработано для их идентификации [86]. Однако, несмотря на известные проблемы с точностью, *ab initio* методы предсказания генов позволяют получить первичный полногеномный набор генных моделей, которые впоследствии можно уточнить и верифицировать с помощью сравнительных и экспериментальных методов.

1.3.3 Использование данных РНК-секвенирования и белковых баз данных для повышения точности предсказания генов

Секвенирование полноразмерных мРНК или их фрагментов, представляющих экспрессирующиеся участки гена (expressed sequence tags, ESTs), а также наличие базы известных белков вида, чей геном аннотируется, являются основным подходом к повышению точности *ab initio* аннотации генов. Вместо того, чтобы полагаться на статистические свойства кодирующих последовательностей и сигналы, которые их определяют, прямое выравнивание транскрибируемых или транслируемых из генома последовательностей дает возможность опираться на

экспериментальные доказательства структуры гена. Проблема выравнивания таких доказательств на «голый» геном была решена с помощью алгоритмов динамического программирования, которые учитывают сайты сплайсинга и экзон-интронную структуру эукариотических генов [34]. К классическим инструментам для выравнивания транскриптов на геномную последовательность относят TopHat [87], STAR [88] и Hisat2 [89]. Процесс выравнивания аминокислотных последовательностей на геном требует трансляции геномной последовательности в несколько вариантов с разной рамкой считывания и стрэнда ДНК, и нахождения оптимального выравнивания мультиэкзонной структуры с родственным белком. Выравнивание аминокислотных последовательностей для предсказания генов в больших геномах может быть чрезвычайно время- и ресурсоемким вычислительным процессом. Однако использование РНК и белковых данных способны существенно повысить точность предсказания.

В основе сборки всегда лежат риды, сгенерированные в процессе секвенирования, и, хотя современные секвенаторы имеют относительно низкий уровень ошибок, данные, которые они производят, не свободны от ошибок [90]. Такие свойства последовательности как относительное содержание гуаниновых и цитозиновых нуклеотидов, длина, состав *k*-меров напрямую влияют на качество сборки. Первым шагом в *de novo* сборке транскриптома является контроль качества, включающий в себя фильтрацию и обрезку прочтений по уровню качества и удаление адаптеров. Данные секвенирования часто содержат неопределенные нуклеотиды (обозначаемые символом N в последовательности). Участки, содержащие большое количество таких неопределенных нуклеотидов, рекомендуется удалять, так как их наличие мешает корректной сборке и последующему анализу. Точно так же данные могут быть отфильтрованы, чтобы сохранить только те прочтения или их части, которые содержат основания с наилучшим показателем качества (Q). Этот показатель [91] отражает вероятность того, что конкретный нуклеотид был определен верно. Прочтения, содержащие некоторое предельное количество нуклеотидов с низким показателем качества, могут либо полностью исключаться, либо обрезаться, если падение качества

наблюдается на концах прочтений, что часто происходит при секвенировании, например, Illumina. Так же зачастую исключаются очень короткие прочтения (<30 п.н.) [92].

Современные эксперименты по РНК-секвенированию генерируют сотни миллионов прочтений с целью реконструкции всех экспрессирующихся транскриптов для получения референсных транскриптомов. Хотя это повышает вероятность обнаружения транскриптов с низкой экспрессией, такой подход также производит слишком много прочтений уже достаточно хорошо представленных транскриптов, создавая переизбыток прочтений для определенных генов, что ведет к увеличению времени обработки и ресурсоемкости при ассемблировании. Чтобы избежать этого, применяется нормализация, при которой прочтения количественно оцениваются на основе содержания в них *k*-меров и либо сохраняются, либо удаляются в соответствии с определенными пользователем пороговыми значениями. Результатом является существенное сокращение объема прочтений таким образом, что может быть достигнута полная реконструкция подавляющего большинства транскриптов, вместе с уменьшением количества прочтений, поступающих в ассемблер [92].

Для *de novo* сборки транскриптомов доступно большое количество ассемблеров, наиболее популярные из которых — это Trinity [93], SOAPdenovo-Trans [94], и rnaSPAdes [95]. Trinity и rnaSPAdes применяют стратегию множественных *k*-меров, стремясь использовать преимущества малых и больших длин *k*-меров для максимального восстановления транскриптов [92].

Немаловажным этапом является проверка качества и полноты полученной сборки. В идеальном случае транскриптомная сборка должна восстановить большую часть секвенированных транскриптов. Одним из общепринятых методов для оценки полноты сборки является проверка на наличие в ней ортологов универсальных, постоянно экспрессирующихся генов, которые встречаются в геномах всех представителей определённой группы (позвоночные, членистоногие, высшие растения и т.д.). Такой анализ может быть выполнен с использованием программы BUSCO (Benchmarking Universal Single-Copy Orthologs) [96], которая

использует обновляемые курируемые наборы универсальных однокопийных генов из базы данных ортологов OrthoDB [97]. Полнота сборки в данном случае оценивается из того, сколько универсальных генов имеют совпадения во проверяемом транскриптоме и являются ли эти совпадения дублированными, фрагментированными или полноразмерными [92].

Использование РНК-данных имеет несколько важных оговорок. РНК-секвенирование не охватывает все гены в геноме, некоторые гены экспрессируются на низком уровне или только в нескольких тканях, на некоторых стадиях онтогенеза или в определённых условиях, и они могут быть полностью упущены. Кроме того, многие из экспрессируемых транскриптов представляют собой не полностью сплайсированные транскрипты или могут быть просто артефактом. Следовательно, необходима независимая проверка, прежде чем присвоить любой экспрессируемой области статус функционального гена. Даже для генов, которые показывают неоднократную высокую экспрессию, определение того, кодируют ли они белки или представляют собой некодирующие РНК, остается часто трудно решаемой проблемой [64].

1.3.4 Функциональная аннотация на основе гомологии

После нахождения генов и определения их структуры, следующим необходимым шагом является присвоение генам биологической функции. Выполнение этого этапа все еще остается сложной задачей, несмотря на накопление обширных знаний о функции генов в модельных и культивируемых видах. До сих пор существует большой процент обнаруженных у разных видов генов, функция которых не установлена [98]. Среди всех растений *A. thaliana* по сей день остается эталонным, модельным видом с наиболее изученным геномом и тщательно проаннотированными генами — во многом благодаря огромным усилиям ресурсов The Arabidopsis Information Resource (TAIR) [48] и AraPort [41,99], которые объединяют аннотации, курируемые сообществом с информацией, доступной из литературных данных. Тем не менее, несмотря на такой обширный ресурс и более чем двадцать лет, прошедших после публикации первой последовательности генома арабидопсиса в 2000 году [100], текущее состояние

сборки арабидопсиса все еще «почти полное», так как в геноме все еще остаются неразрешенные пробелы, предположительно состоящие из повторяющихся последовательностей — теломерных и центромерных участков, кластеров 5S рДНК и областей ядрышковых организаторов, содержащих 45S рДНК [98,101].

Учитывая текущий уровень аннотирования растительных геномов, неудивительно, что часто единственным практичным методом аннотирования становится поиск сходства последовательностей с модельным *A. thaliana*. На практике, часто выполняется простой поиск BLAST с использованием генома/транскриптома в качестве входных данных и протеома *A. thaliana* в качестве референсной базы. Другой подход заключаются в том, чтобы использовать поиск открытых рамок считывания, кодирующих функциональные домены, с помощью TransDecoder (https://github.com/TransDecoder). Инструменты, которые специализируются на идентификации доменов в последовательности, имеют преимущества перед простым поиском по сходству, поскольку последовательности доменов обычно высоко консервативны между генами. Домены часто представлены в виде профилей скрытых марковских моделей (HMM), полученных из множественных выравниваний последовательностей, происходящих от нескольких видов — таким образом фиксируется типичное разнообразие последовательностей в отдельных сайтах. Это обеспечивает более чувствительный подход к проблеме аннотации последовательности [98].

Функциональная аннотация обычно подразумевает перенос информации о функции от одного гена к другому, при этом предполагается, что первоначальная функциональная аннотация верна. Однако, даже в тех случаях, когда перенос аннотации был успешен, необходимо решать вопрос о ее качестве. Учитывая, что белки обычно состоят из одного или нескольких отдельных консервативных доменов, встроенных в более общие кодирующие участки, методы, учитывающие только сходство последовательностей, но не учитывающие, что для выполнения функции необходимы определенные домены, могут привести к неправильной аннотации. Следует иметь в виду, что к функциональным аннотациям следует

относиться с осторожностью и рассматривать их скорее как рабочие гипотезы, которые могут нуждаться в экспериментальной проверке [98].

Наиболее широко используемой базой для функциональной аннотации является «Gene Ontology» (GO) — онтология, которая классифицирует генные продукты тремя отдельными категориями («терминами»): «биологический процесс», «клеточный компонент» и «молекулярная функция». Онтология GO структурирована как ориентированный ациклический граф, позволяющий вывести более общие термины из конкретного, что позволяет дополнительно группировать данные [102,103]. Существует ряд программных решений, выполняющих присвоение функций на основе популярных онтологий, такие как KEGG Automatic Annotation Server (KAAS) [104], Mercator [105], AHRD (https://github.com/groupschoof/AHRD) или Blast2GO [106,107]. Показано, что последний дает самую высокую точность аннотирования, но имеет самое больше время работы [98].

1.4 Структура промоторной области и типичные регуляторные мотивы в геноме растений

Транскрипция — это процесс передачи информации от гена к матричной РНК, который осуществляется РНК-полимеразой II. Регуляция этого процесса достигается благодаря связыванию факторов транскрипции (ТФ) с геномными сайтами, содержащими регуляторные нуклеотидные мотивы, которые находятся в пределах промоторной области, 1000 п.н. выше (левее) от сайтов начала транскрипции (transcription start site, TSS). Положение TSS соответствует первому нуклеотиду, транскрибируемому РНК-полимеразой II. Эукариотические гены могут иметь несколько альтернативных TSS [108,109].

Базовый промотор представляет собой участок ДНК длиной до 250 п.н., расположенный непосредственно перед TSS и необходимый для инициации транскрипции. Существует два типа инициации транскрипции: сфокусированная, обычно связанная с регуляцией тканеспецифичных генов и генов ответа на стресс, и диспергированная, обычно встречающаяся в генах домашнего хозяйства при

конститутивном паттерне экспрессии [110]. Если область инициации транскрипции широкая, то соответствующий диспергированной транскрипции участок называется областью начала транскрипции (transcription start region, TSR). В ключевых областях промотора наиболее известным регуляторным мотивом является TATA-бокс (консенсус TATA(A/T)A(A/T)), который распознается TATA-связывающим белком и встречается в до 60% всех промоторов [109,111–115]. Еще одним распространенным мотивом является Inr с консенсусной последовательностью YYA+1NT/AYY, который обнаруживается в месте начала транскрипции. Inr мотив более широко распространен, чем любой другой мотив экспрессии [110], и обычно встречается в генах домашнего хозяйства, транскрипция которых инициируется сверхдиспергированными промоторными областями [116], в то время как промоторы, содержащие мотив TATA, обычно более узкие и связаны с экспрессией генов, зависящей от типа ткани или контекста [117].

Поиск промоторов является важным этапом аннотации генома, так как они играют ключевую роль в регуляции транскрипции генов [114]. В настоящее время существуют высокопроизводительные методы идентификации сайтов начала транскрипции и связывания транскрипционных факторов, которые позволяют получать значительное количество данных о регуляторных элементах растений. Такие методы, как иммунопреципитация хроматина в сочетании с анализом на микрочипах или секвенированием (ChIP-chip и ChIP-seq), определение гиперчувствительности к ДНКазе I (DNase I-hypersensitive sites, DHS) и кэп-анализ экспрессии генов (cap analysis of gene expression, CAGE) позволили накопить значительный объем данных о регуляторных элементах растений [118,119]. Однако, эти методы являются трудоемкими и затратными. Поэтому были разработаны вычислительные подходы, которые позволяют быстро и точно прогнозировать местоположение TSS и регуляторных мотивов в масштабе всего генома. Подходы полногеномного обнаружения новых цис-регуляторных мотивов с использованием позиционно-весовых матриц (PWM) и данных экспрессии были успешно реализованы для риса и арабидопсиса [115,120], хмеля [121] и

виноградной лозы [122]. Полногеномный анализ основных промоторных элементов с использованием PWM и прогнозирования на основе ортологов был выполнен для нескольких видов однодольных и двудольных растений [123].

Форма молекулы ДНК определяется «кривизной» («искривленностью») и «изгибаемостью» (гибкостью или «мягкостью»), которые обусловлены внутренней энергией системы или воздействием внешней силы. Набор форм, которые ДНК принимает без воздействия внешних сил называется «кривизной» ДНК, тогда как способность ДНК деформироваться под действием внешней силы, называется «изгибаемостью» [124]. Оба эти параметра зависят от состава последовательности ДНК. Промоторы отличаются от других участков генома своей низкой стабильностью ДНК, высокой изгибаемостью и кривизной [113,125–127]. Для определения положения TSS используется метод изменения стандартной свободной энергии ДНК дуплекса, который успешно применяется для идентификации промоторов у разных видов эукариот [128]. Изгибаемость ДНК в области перед TSS также является важной особенностью, поскольку она взаимодействует с ДНК-связывающими белками [126,129–131]. Кроме этого, промоторы также обладают GC-асимметрией и отличаются пониженной генетической изменчивостью. [113,127,132]. Для нескольких видов растений и животных был отмечен избыток цитозиновых нуклеотидов над гуаниновыми (GC-skew) в смысловом стрэнде ДНК вокруг TSS [133–135]. Было высказано предположение, что GC-skew вокруг TSS может быть связан с более высокой вероятностью дезаминирования цитозина во время транскрипции из-за предпочтительной защиты РНК-полимеразой нуклеотидов на нетранскрибируемой цепи [135].

Исследования показали, что свойства кодирующей и промоторной областей имеют взаимосвязь. Важным свойством кодирующих областей является частота встречаемости гуаниновых и цитозиновых нуклеотидов в третьем положении (GC3). Это связано с тем, что нуклеотиды в третьем положении менее подвержены отбору, чем первые два, из-за вырожденности генетического кода. Было замечено, что на основе GC_3-состава геномы можно разделить на две группы, с

унимодальным и бимодальным распределением GC_3. Так, большинство секвенированных в настоящее время геномов трав имеют бимодальное распределение GC_3, в то время как кодирующие последовательности двудольных растений показывают унимодальное распределение [136,137]. Ранее считалось, что бимодальность GC_3 является специфической особенностью геномов трав. Позже было показано, что GC_3-состав у других видов однодольных, таких как куркума длинная (*Curcuma longa* L.), имбирь лекарственный *(Zingiber officinale* Roscoe), масличная пальма *(Elaeis guineensis* Jacq.) и зантедеския (*Zantedeschia aethiopica* (L.) Spreng.), также демонстрирует бимодальное распределение GC_3 [138–140]. Было показано, что гены с более высоким содержанием GC_3 также имеют более высокую частоту встречаемости TATA-бокса и с большей вероятностью связаны со стрессом [137].

1.5 Применение микросателлитных маркеров для изучения генетического разнообразия растений

Высокоповторяющиеся последовательности низкой сложности занимают большую долю в геномах растений. Одним из наиболее частых классов таких последовательностей являются микросателлитные повторы (simple sequence repeats, SSRs), определяемые как многократное повторение нуклеотидного мотива размером 1–9 п.н. Большинство микросателлитных повторов находятся в некодирующих участках генома [141] и обладают высокой скоростью мутирования, в основном за счет потери или добавления «мономеров», вызванных проскальзыванием или сбоем работы ДНК-полимеразы во время репликации [142]. Микросателлитные локусы наследуются кодоминантно и имеют равномерное распределение по хромосомам. Все это позволяет использовать их для оценки внутри- и межпопуляционного полиморфизма с высокой точностью [143]. Еще одно преимущество микросателлитного анализа заключается в его относительной простоте и экономической выгоде. SSR маркеры требуют лишь проведения ПЦР реакции с известными праймерами; визуализацию продуктов реакции и генотипирование можно осуществить с помощью гель-электрофореза, обычно в

полиакриламидном геле. Это позволяет применять микросателлитный анализ для маркерной селекции и анализа генетического разнообразия практически в любых лабораториях, с минимальным оснащением. Было показано успешное применение микросателлитных маркеров для индивидуального и популяционно-генетического анализа, филогенетического анализа близкородственных видов внутри рода, а также для определения происхождения деревьев [144,145].

Одним из недостатков микросателлитных маркеров до недавнего времени считались трудоёмкость и дороговизна в их разработке, требовавшей получения геномных библиотек, обогащенных микросателлитными локусами с последующим их клонированием и секвенированием, что осложняло их использование для исследования новых видов. Польза SSR, как генетических маркеров для древесных видов также была ограниченна тем, что разработанные праймеры часто не могли амплифицировать тот же самый локус у родственных таксонов, если данный повтор не фланкирован высоко-консервативными последовательностями. Еще одна проблема, связанная с SSR-маркерами это присутствие «нуль» аллелей — отсутствие амплификации ввиду неполной комплементарности праймеров из-за наличия мутаций в данном регионе. Такое чаще происходит при использовании праймеров, изначально разработанных для другого родственного вида. Наличие нуль-аллелей ведет к сложностям в оценке частот аллелей и недооценке уровня гетерозиготности [146]. Отчасти это может компенсироваться выявлением и анализом нуль-аллелей с помощью специализированных программ, таких как Micro-Checker [147].

Для изучения популяций применяют не только ядерные микросателлитные маркеры, но и цитоплазматические: митохондриальные [141] и хлоропластные [148,149]. Несмотря на то, что средний уровень их изменчивости ниже, чем у ядерных, цитоплазматические маркеры представляют особый интерес ввиду специфического характера наследования у разных групп растений. У большинства видов передача как хлоропластного, так и митохондриального геномов осуществляется по материнской линии. Однако, например, у хвойных геном пластид передается потомству с пыльцой, а геном митохондрий — с семенами

[150,151], что позволяет изучать наследование по разным линиям — материнской и отцовской, соответственно.

Для представителей рода *Larix* за последнее время были опубликованы данные по разработке микросателлитных маркеров для лиственниц японской [152], европейской [153], Гмелина [152], Лайэля и западной [154,155]. В. Л. Семериков вместе с коллегами изучали генетическое разнообразие лиственницы сибирской на Урале, используя полиморфизм цитоплазматических маркеров и частично ядерных генов [25,156,157]. Н.В. Орешкова с соавторами в 2013 году впервые применили для оценки генетического разнообразия популяций лиственницы сибирской набор из семи микросателлитных маркеров, разработанных изначально для других видов лиственницы [28]. Авторы уточняют, что полученные высокие значения наблюдаемой гомозиготности, по сравнению с ожидаемой, связаны не только с инбридингом в анализируемых популяциях, но скорее с присутствием скрытых нуль-аллелей, которые могут возникать из-за недостаточной комплиментарности используемых праймеров, предназначенных для работы с другими видами *Larix*. Это приводит к недоамплификации аллелей, по которым праймер не полностью совпадает с сайтом отжига, и как результат к неправильному генотипированию гетерозигот как гомозигот и искусственному завышению их частоты, создавая при этом ложную видимость инбридинга. В другом исследовании для оценки генетического полиморфизма и индивидуальной гетерозиготности лиственницы сибирской использовались восемь полиморфных ядерных микросателлитных локусов, также разработанных для других видов лиственницы [27]. Аналогичным образом авторы отмечают, что малое количество полиморфных маркеров потенциально способно привести к уменьшению информативности, и, как следствие, к искаженной оценке генетического разнообразия.

1.5.1 Методика разработки и анализа микросателлитных маркеров

Общая процедура создания нового набора полиморфных SSR-маркеров для конкретного вида заключается в:

– поиске в нуклеотидных последовательностях подходящих локусов с достаточным числом тандемных повторов, имеющих достаточно длинные

фланкирующие последовательности для выбора сайтов отжига и дизайна ПЦР праймеров,

– подборе соответствующих праймеров,

– оптимизации условий ПЦР на относительно небольшом числе образцов ДНК,

– разделении амплифицированных фрагментов с помощью гель-электрофореза,

– тестировании праймеров на большом количестве образцов ДНК.

Зачастую разработанные маркеры тестируют так же на близкородственных видах, с целью проверить их универсальность и полиморфность. Для конкретного вида могут быть адаптированы микросателлитные локусы известные для другого, близкородственного вида или рода. Известно, что фланкирующие участки имеют более низкую скорость мутации, чем сами SSR-повторы, что делает возможным кросс-видовое применение некоторых праймеров [158], однако в этом случае есть опасность неполного соответствия праймеров сайтам отжига, что будет приводить к появлению нуль-аллелей и ошибочному генотипированию гетерозигот по таким аллелями как гомозигот.

1.5.2 Идентификация тандемных повторов и подбор праймеров для микросателлитных локусов

Поиск тандемных повторов может быть осуществлен при помощи разнообразного программного обеспечения, такого как Tandem Repeat Finder (TRF) [159] и MISA [160], предназначенного для поиска тандемных повторов. Для идентификации микросателлитных повторов в геномных данных большого объёма было разработано программное обеспечение GMATo [161].

Для проведения успешной амплификации целевых маркерных последовательностей важно, чтобы праймеры к ним были специфичны — то есть последовательности праймеров для данного маркера должны встречаться в геноме только в одном месте. Для этого при помощи инструментов выравнивания необходимо исключить последовательности праймеров, встречающиеся в геноме более одного раза.

Следующим шагом на пути к созданию SSR-маркеров является подбор

праймеров и их проверка с помощью ПЦР амплификации. Длина прямого и обратного праймеров может варьировать от 18 до 30 п.н. Однако, длина в 21–23 п.н. считается наиболее оптимальной для амплификации микросателлитных локусов [162]. Желательно, чтобы праймеры имели 40–60% GC-состав. Образование димеров праймеров можно избежать путем поиска более подходящих сайтов отжига, праймеры для которых не образуют димеров, или с использованием специальной ДНК полимеразы (hot start *Taq* DNA polymerase) и «горячего» старта в протоколе ПЦР.

1.5.3 Оптимизация условий ПЦР и гель-электрофореза для анализа микросателлитных маркеров

При проведении SSR-анализа ПЦР используется для амплифицирования микросателлитных локусов с помощью подобранных праймеров комплиментарных сайтам отжига фланкирующим микросателлитные локусы, с последующим определением длин фрагментов. Обычно ПЦР программа состоит из примерно 35 циклов смены температурного режима, каждый из которых включает в себя стадию денатурации при 95°C, отжига праймеров при 55–65°C и элонгации при 72°C. Для успешной амплификации маркерных локусов необходимо соответствующим образом оптимизировать условия ПЦР, в противном случае велика вероятность различных осложнений, таких как отсутствие амплификации, низкий выход целевого продукта и большое количество неспецифичных фрагментов, возникающих вследствие ошибок в связывании праймеров и образовании праймер-димеров. Подбор правильной температуры отжига праймеров играет здесь крайне важную роль. При слишком высокой температуре не произойдет связывание праймера с ДНК-матрицей, а при слишком низкой может происходить неспецифическое связывание. Слишком большое количество циклов также ведет к появлению неспецифичных продуктов.

Другая разновидность ПЦР-программ — так называемый «горячий старт», который используется для предотвращения образования димеров, самоотжига и самоудлиннения через образование «шпилек», неспецифического отжига праймеров на первых стадиях ПЦР. С неспецифической амплификацией так же

можно бороться посредством программ с «touchdown». В этом случае целевые локусы амплифицируются при высокой температуре отжига первые несколько циклов реакции, с последующим понижением температуры в последующих 10–15 циклах до определённого уровня.

Для генотипирования SSR-локусов на основании длин амплифицированных фрагментов используются три типа гель-электрофореза. Самый простой из них — электрофорез в агарозном геле высокого разрешения с окрашиванием в растворе бромистого этидия [163]. Второй метод использует полиакриламидный гель, который, хотя и является более дорогим и сложным в приготовлении, однако имеет еще более высокую разрешающую способность, чем агарозный. Наконец, самым точным методом для определения полиморфизма длин микросателлитных локусов является капиллярный электрофорез с использованием секвенаторов типа ABI PRISM (Applied Biosystems). Данная методика показывает крайне высокую эффективность, а также позволяет создавать мультиплексные панели, уменьшая стоимость и трудоемкость анализа. Минусом является его высокая стоимость, и недоступность для некоторых лабораторий, ввиду отсутствия специализированного оборудования.

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1 Аннотация ядерного генома лиственницы сибирской

Для аннотации была использована сборка генома лиственницы сибирской v1.0 (NCBI GCA_004151065.1) общей длиной 12,34 млрд. п.н. и минимальной длиной контига 200 п.н. (Таблица 2), полученная в лаборатории лесной геномики СФУ [12].

Таблица 2. Статистика сборки генома лиственницы сибирской.

Сборка	Количество, млн	N50, п.н.	Максимальная длина, п.н.	Суммарная длина, млрд. п.н
Контиги	12,40	1 074	128 642	7,99
Скаффолды	11,33	6 443	354 326	12,34

Для обеспечения поддержкой предсказанных генных моделей были использованы референсные транскриптомы пяти тканей — почки, хвои, камбия, проростка и побега первого года, также полученные в лаборатории лесной геномики СФУ. Прочтения транскриптомов доступны в NCBI SRA: SRX9464971, SRX14986114, SRX14997110, SRX14997111 и SRX14997112. Сборки транскриптомов доступны в NCBI: GIXH00000000, GJYD00000000, GJYL00000000, GJYN00000000 и GJYW00000000 [20].

2.1.1 Анализ и маскирование высокоповторяющихся элементов

Для поиска тандемных повторов и мобильных элементов был использован RepeatModeler v.1.0.11 [164], использующий программы обнаружения повторов *de novo* RepeatScout и RECON [72,74]. Поскольку для оптимизации компьютерного времени RepeatScout использует для анализа не все скаффолды или контиги, а только некоторую случайно выбранную их часть, было решено взять для анализа только скаффолды длиной более 100 Кб (2869 скаффолдов суммарной длиной 360 млн. п.н.). RepeatMasker open-4.0.6 [165] использовался для маскирования областей низкой сложности и повторов на основе комбинированной библиотеки повторов; база RepBase RepeatMasker 2017.01.27 [166] была расширена видоспецифичной библиотекой повторов для лиственницы, полученной с помощью RepeatModeler open-1.0.8, который был запущен с настройками по умолчанию. Эта

комбинированная база данных использовалась в дальнейшем в пайплайне MAKER2 для маскировки повторов (**Рисунок 3**) [20].

Рисунок 3. Схема маскировки повторов в геномной сборке

Для оценки относительной представленности известных семейств повторов был использован RepeatMasker на полной сборке генома (12,34 млрд. п.н.). Для получения наиболее полной библиотеки повторов, библиотека *de novo*, созданная RepeatModeler, была дополнена кластеризацией часто встречающихся ридов из данных полногеномного секвенирования. Кластеры прочтений были собраны с помощью Inchworm из TrinityRnaSeq v2.2.0, в результате чего были получены консенсусные последовательности, которые должны представлять сильно повторяющиеся участки генома лиственницы. Нераспознанные элементы из библиотеки повторов *de novo*, созданной RepeatModeler, и кластеры часто встречающихся прочтений были охарактеризованы с помощью TEclass v2.1.3, который классифицирует транспозоны с использованием метода опорных векторов и нейронной сети LVQ [167] (Рисунок 4).

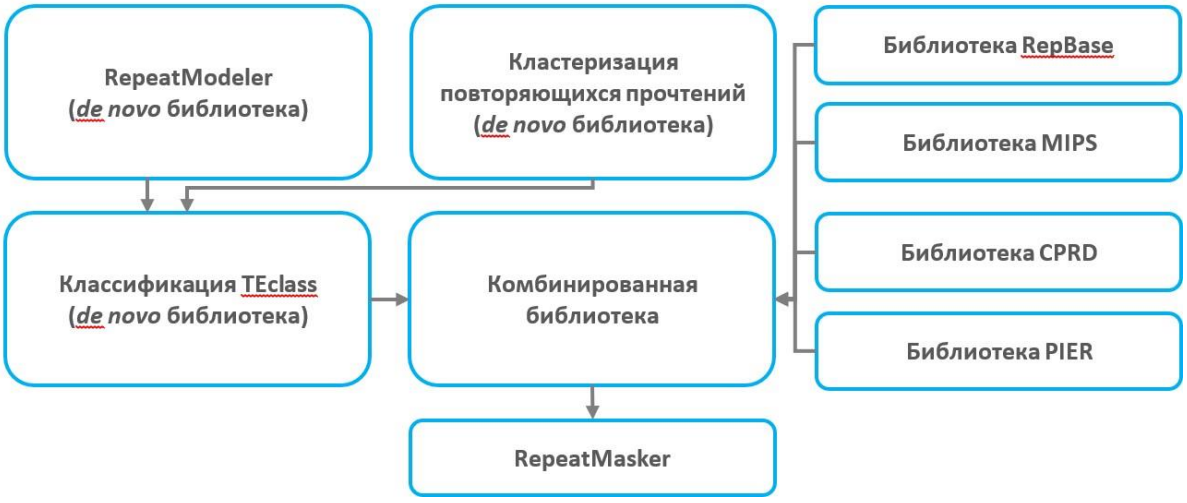


Рисунок 4. Схема анализа представленности классов повторов в геномной сборке

Для поиска мобильных элементов и оценки их относительной представленности в геноме была использована комбинированная база повторов,

включающая библиотеку RepeatModeler, классифицированную с помощью TEclass, библиотеку RepBase (версия 2017.01.27), библиотеку базы данных повторяющихся элементов MIPS [168], пользовательскую базу данных повторяющихся элементов CPRD [169] и библиотеку диспергированных повторов сосны PIER v1.0 [21,169] (**Ошибка! Источник ссылки не найден.**Рисунок 4). Часть длинных прочтений O xford Nanopore, доступная для лиственницы сибирской, также использовалась для оценки доли повторов в геноме. Пользовательский R-скрипт применялся для анализа результатов RepeatMasker в соответствии с классификацией RepBase. Классификация мобильных элементов проводилась в соответствии с обновлением Repbase [170]. Для поиска тандемных повторов были использованы программы GMATo [161] и TRF [20,159].

2.1.2 Оценка времени вставки ретротранспозонов LTR-RT

Существует два распространенных метода оценки времени вставки ретротранспозонов с длинными концевыми повторами (LTR-RT):

1) измерение расхождения последовательностей между двумя фланкирующими LTR и расчёт времени их расхождения с учётом скорости видоспецифичных мутаций;

2) анализ попарных генетических расстояний между последовательностями, кодирующими ретротранспозон, которые принадлежат к паралогичным элементам одной и той же монофилетической группы.

Хотя эти методы могут давать разные оценки времени для некоторых линий LTR, профили распределения времени, полученные обоими методами, аналогичны [171].

Для дополнительного *de novo* поиска элементов LTR-RT был использован LTRharvest [78] с параметрами «-tis -suf -lcp -des -ssp -sds -dna». Для исключения потенциальных ложноположительных предсказаний и получения более строгого набора LTR-RT, был использован LTR_retriever [172], обрабатывающий результаты LTRharvest, с параметрами «-u 1,57e-8 -missrate 0,4 -noanno» [20].

Zhou с соавторами [173] создали библиотеку LTR элементов для 301 растения, включая доступные полные геномы 10 голосеменных растений. Интактные

длинно-концевые повторы (367 повторов) из библиотеки Zhou с соавторами [173] были использованы для сравнения с повторами, найденными с помощью LTRharvest и LTR_retriever, при помощи blastn. Последовательности без совпадений были добавлены в окончательную базу длинно-концевых повторов для лиственницы сибирской. Оценка времени вставки других видов голосеменных была проведена на повторах также взятых из [173]. Расхождение последовательностей было рассчитано с использованием модели замен Jukes-Cantor [172]:

$$T = \frac{d}{2\mu}, \quad d = \frac{-3}{4} ln \left(1 - \frac{4}{3}p\right), \qquad (1)$$

где d — генетическое расстояние по Jukes-Cantor (показатель степени расхождения), μ — частота мутаций и p — доля различий последовательностей (p = 1 — идентичность, которая аппроксимируется с помощью blastn). Время вставки, преобразовано в миллионы лет, при скорости синонимичных замен $\mu = 1{,}57 \times 10^{-8}$ на сайт в год [174].

2.1.3 Идентификация генов с лейцин богатыми повторами (LRR)

Поиск LRR проводился в ORF транскриптов лиственницы сибирской (NCBI SRA SRX9464971, SRX14986114, SRX14997110, SRX14997111 и SRX14997112). ORF были идентифицированы с помощью Transdecoder v.5.5.0 (https://github.com/TransDecoder). ORF транскриптов были просканированы с помощью HMMER 3.2.1 [175] с использованием моделей Pfam [176]. Дополнительно был проведен поиск R-генов NBS (NB-ARC PF00931из базы Pfam 32.0) для проверки принадлежности некоторых последовательностей с LRR к R-генам [20].

2.1.4 Тренировка программы AUGUSTUS для *de novo* поиска генных моделей

Для автоматизированной генной аннотации сборки генома был использован MAKER2 [177]. MAKER2 поддерживает несколько предикторов генов *ab initio*, включая SNAP [178], GeneMark [179] и AUGUSTUS [180]. Все они требуют предварительного обучения для получения видоспецифичных параметров,

описывающих паттерны экзон-интронной структуры. В данной работе для поиска предварительных генных моделей был выполнен с помощью AUGUSTUS v3.2.1 [180,181].

Обучение AUGUSTUS проводилось в несколько этапов. В начале предиктор запускался с параметрами по умолчанию для *Arabidopsis thaliana*, в результате чего было получено 916 тыс. предварительных генных моделей с относительно низкой точностью предсказания. Чтобы собрать начальный обучающий набор генов, были использованы данные секвенирования РНК. Прочтения транскриптома были сопоставлены с геномом с помощью TopHat [87], полученные выравнивания были использованы с Cufflinks для аннотирования кодирующих областей [182]. Это позволило получить 16 тысяч генных моделей, предсказанный на основе транскриптома. Затем два набора генов от AUGUSTUS и Cufflinks были объединены вместе, чтобы отфильтровать исходный набор предсказаний, и выбрать гены с более чем одним экзоном. Этот процесс запуска AUGUSTUS с набором генов, полученным из транскриптома, фильтрации полученных на выходе предсказаний и получения улучшенных параметров обучения повторялся три раза (Ошибка! Источник ссылки не найден.), пока точность предсказания не стала с равнимой со средней точностью для AUGUSTUS. Окончательные параметры обучения использовались в MAKER2 [20].

Рисунок 5. Схема аннотации генома лиственницы сибирской

2.1.5 Аннотация с использованием программы MAKER2

MAKER2 версии 2.31.8 использовался для получения структурной аннотации. Для BLAST использовалась версия ncbi-blast-2.2.29+. В качестве вспомогательных данных использовались транскриптом лиственницы сибирской и общедоступные сборки транскриптомов родственных видов хвойных, депонированные на веб-сайте проекта PlantGenIE (https://plantgenie.org). В качестве референсной базы данных белков использовалась база Uniprot (https://www.uniprot.org/) [20].

Аннотирование генома с помощью MAKER2 было выполнено на 448 ядерном вычислительном кластере СФУ с 56 серверами IBM BladeCenter HS21 (16 ГБ ОЗУ на сервер). Организация запуска MAKER2 на суперкомпьютерном кластере проводилась совместно с сотрудниками кафедры высокопроизводительных вычислений СФУ под руководством к.т.н. Кузьмина Д.А. и к.ф.-м.н. Феранчука С.И.

2.1.6 Оценка полноты сборки и функциональная аннотация

Оценка полноты проводилась с помощью BUSCO v4.0.5 [96] на основе референсной базы Embryophyta и аминокислотных последовательностей, полученных из аннотации MAKER2 для генома лиственницы сибирской. Наборы белков для других видов голосеменных были взяты из базы данных treegenes (https://treegenesdb.org).

Поскольку в референсной базе меньшего размера вероятность определения отдаленных гомологов выше [183], из базы последовательностей NCBI nr была сформирована выборка, путем фильтрации по идентификатору таксономии на уровне Embryophyta. Поиск белковых доменов проводился с помощью InterProScan [184,185] на веб-сервере EMBL-EBI. Картирование терминов генной онтологии (GO) было выполнено с использованием программного обеспечения Blast2GO, интегрированного в платформу OmixBox [106,107]. Все предсказанные гены были картированы против базы данных NCBI GenBank nr с использованием BLASTp. Совпадения с бактериями, грибами и археями (e-value < 1×10^{-5}, процент совпадений

> 20, доля покрытых high-scoring pair > 20) были удалены, чтобы исключить гены, которые потенциально могут представлять белки других организмов [20].

Для сравнения лиственницы с другими голосеменными и покрытосеменными видами, аннотации геномов пяти других хвойных (*P. glauca, P. abies, P. lambertiana, P. taeda* и *P. menziesii)* и пяти покрытосеменных растений (*Betula pendula, Fagus sylvatica* L., *Populus trichocarpa* Torr. & A.Gray ex. Hook., *Quercus robur* L. и *Vitis vinifera* L.*)* использовались для поиска ассоциаций с терминами генной онтологии. Для выявления терминов GO со значимым различием в числе картированных генов использовался тест пропорций. Для коррекции *p-value* были использованы два метода расчёта FDR согласно [186] и [187].

2.2 Аннотация органельных геномов лиственницы сибирской

2.2.1 Сборка и аннотация хлоропластного генома

Для получения сборки генома хлоропласта были использованы данные полногеномного секвенирования трех деревьев лиственницы сибирской, полученные с помощью секвенирующей платформы Illumina HiSeq 2000 [12]. Образцы ДНК были выделены из хвои и гаплоидного каллуса трех деревьев лиственницы сибирской, представляющих разные регионы России — Урал, Красноярский край и Республику Хакасия. Для секвенирования были использованы библиотеки paired-end (PE) и mate-pair (MP), с размерами фрагментов 400–500 п.н. (Уральское и Красноярское деревья) и 300–400 п.н. (Хакасское дерево). В качестве референса для сборки и аннотации были использованы хлоропластные геномы *Larix decidua* Mill. [188] и *L. occidentalis* Nutt. [189] (NCBI Genbank AB501189.1 и FJ899578.1, соответственно). Полученные прочтения были картированы на референсные геномы хлоропластов *L. decidua* и *L. occidentalis* с помощью Bowtie2 [190]. Выровненные прочтения были собраны с помощью ассемблера SPAdes [191]. Полученные контиги снова картировались на референсный геном *L. decidua* с помощью BLAST, таким образом были отобраны контиги, представляющие наиболее удачно собранные консервативные участки генома. Эти контиги были использованы в новой итерации сборки SPAdes под

флагом «--trusted-contigs», дающим им боле высокий приоритет [32]. Завершающим этапом сборки был скаффолдинг, который выполнялся с использованием сгенерированных контигов и MP-прочтений с помощью SSPACE [192].

Для аннотации использовался сервис Rapid Annotation with Subsystem Technology (RAST) [193]. Для верификации и уточнения функции гипотетических кодирующих областей, полученную аннотацию сравнивали с аннотациями близкородственных *L. decidua* и *L. occidentalis*, некоторые фрагменты также были проверены вручную с помощью BLAST. Собранный хлоропластный геном *L. sibirica* был депонирован в NCBI GenBank (NC_036811.1) [32].

2.2.2 Сборка и аннотация митохондриального генома

Митохондриальный геном был собран из ДНК, выделенной из хвои референсного дерева лиственницы сибирской, которое использовалось в проекте полногеномного *de novo* секвенирования [12]. Для получения полной сборки было использовано два подхода. В первом случае, общая геномная ДНК была выделена после обогащения мтДНК путем выделения и очистки митохондрий. Выделение интактных митохондрий и получение обогащенной мтДНК проводилось на базе СИФИБР СО РАН под руководством д.б.н. Юрия Михайловича Константинова. Эта ДНК, а также необогащенная высокомолекулярная ДНК, были использованы для секвенирования на платформе HiSeq 2000 (PE и MP библиотеки, 2 ×100 циклов). Во втором случае, тотальная высокомолекулярная ДНК была использована для получения длинных прочтений с использованием MinION с R9 FlowCells (FLO-MIN106, Oxford Nanopore Technologies, Inc., Оксфорд, Великобритания). Секвенирование проводилось сотрудниками лаборатории лесной геномики СФУ [33].

Качество ридов Illumina оценивалось с помощью FastQC v.0.11.5 [194]. Последовательности адаптеров были обрезаны, а короткие прочтения отфильтрованы с минимальным качеством 19 и минимальной длиной 35 п.н. с использованием Trimmomatic v.0.36 [195]. Из PE и MP прочтений была получена предварительная сборка с помощью CLC Assembly Cell v.5.0.0 (QIAGEN

Bioinformatics, Hilden, Germany), BESST [196] и Sealer (https://github.com/bcgsc/abyss/tree/master/Sealer). Длинные прочтения Oxford Nanopore были использованы для улучшения сборки митогенома. Обработка сырых данных и оценка качества прочтений MinION были выполнены с помощью Albacore (https://github.com/dvera/albacore). Гибридная сборка с использованием длинных ридов MinION и коротких ридов PE Illumina проводилась с помощью MaSuRCA v.3.2.8 [197], сотрудниками лаборатории лесной геномики В. Шаровым и Ю.А. Путинцевой. Для извлечения митохондриальных контигов из гибридной сборки использовался поиск BLAST по всем последовательностям митохондриальных растений, доступным в NCBI GenBank. После сопоставления этой сборки с базой данных митохондрий растений было, собрано 9 митохондриальных контигов общей длиной 11,7 млн. п.н. [33]. Оценка точности гибридной сборки была проведена с помощью REAPR v1.0.18 [198].

Для поиска повторов в сборке митогенома был использован RepeatModeler v.1.0.11 [164]. Неклассифицированные повторы были классифицированы с помощью TEclass [167]. В дополнение к библиотекам *de novo* и RepBase [166], для запуска RepeatMasker v. 4.0.6 [165] было использовано несколько дополнительных библиотек повторов: Repeat Element Database [168], Custom Plant Repeat Database [169] и Pine Interspersed Repeats Resource library PIER v1.0 [21]. Анализ результатов RepeatMasker проводился в соответствии с актуальной классификацией RepBase [166] с помощью пользовательского R-скрипта [33].

Митогеном лиственницы сибирской был проверен на гомологию с митогеномами других растений, имевшимися в базе данных NCBI GenBank, с использованием BLAST. Mitofy [199] также использовался для предварительной аннотации митогенома. Гены тРНК были проаннотированы с помощью ARAGORN [200] и tRNAscan-SE [201]. Рибосомальные РНК (рРНК) были проаннотированы с помощью RNAmmer [33,202].

2.3 Предсказание TSS

2.3.1 Геномные данные и фильтрация генов

Сборка генома *Pinus taeda* и аннотация Pita_v2_01 [13,21,55] были взяты из базы данных на https://treegenesdb.org/FTP/Genomes/Pita/v2.01. Сборка генома PG29_v3.0 и соответствующая аннотация для *Picea glauca* [16,203] взяты с ftp://plantgenie.org/Data/ConGenIE/Picea_glauca/PG29/v3.0/. Также была учтена ручная аннотация части генов для сборки PG29_v4.0, которая была добавлена к аннотации версии 3.0. Для *Picea abies* [17] геном Pabies_v1.0 был взят с ftp://plantgenie.org/Data/ConGenIE/Picea_abies/v1.0. Соответствующие аннотации для набора с высокой достоверностью (предсказанные генные модели с гомологией более 70% к референсным белкам) и набора со средней достоверностью (предсказанные генные модели с гомологией от 30% до 70% к референсным белкам), предоставленные авторами, были объединены в один набор. Для *L. sibirica* [12] использовалась сборка генома версии v1.0 (NCBI GCA_004151065.1; [12]) и данные аннотации, описанной в данной работе [204].

Чтобы отфильтровать возможные псевдогены и отобрать наиболее вероятные функциональные гены, все генные модели, извлеченные согласно геномным аннотациям, были сопоставлены с базой данных фрагментов РНК, включая EST и TSA родственных видов с использованием HISAT2 [89], а также с базой растительных белков RefSeq с помощью BLASTp [204].

2.3.2 Предсказание позиций TSS с помощью алгоритма TSSPlant

Предсказание TSS было выполнено на нуклеотидных последовательностях, определенных как области –1000 и +250 п.н. вокруг стартового кодона каждого гена, с использованием программы TSSPlant [205]. TSSPlant использует нейронные сети для оценки таких параметров, как наличие классических растительных промоторных мотивов (TATA, INR, DPE, YP), вариации нуклеотидного состава (CG-skew и AT-skew), оценка представленности олигомеров и др. Поскольку результатом работы является предсказание нескольких возможных сайтов с лучшими показателями, был использован выбор наилучшего предсказания, чтобы

определить один наиболее вероятный TSS для каждого гена. Предполагая, что длина 5'-UTR может быть описана гамма-распределением, был составлен пул длин 5'-UTR на основе аннотаций нескольких модельных растений (*Arabidopsis thaliana, Oryza sativa* L., *Sorghum bicolor* (L.) Moench и *Populus trichocarpa*) и вычислены параметры $\boldsymbol{k}$ и $\boldsymbol{\theta}$, определяющие форму и масштаб распределения длин 5'-UTR [204]*:*

$$\Theta = v/m,\ k = m/\theta \qquad (2)$$

Для выбора наиболее вероятных положений TSS была применена функция плотности вероятности [204].

2.3.3 Анализ нуклеотидного состава промоторов и генов

Анализ частот нуклеотидов в промоторах был проведен на последовательностях, центрированных по позиции предсказанного TSS (–1000, +1000 вокруг TSS). Расчет частот мотивов CA и TATA выполнен в скользящем окне (ширина = 20 п.н., шаг = 10 п.н.) с использованием пакета stringr R. CG-skew определялся как пропорция (C-G)/(C+G) и рассчитывался в скользящем окне (ширина = 50 п.н., шаг = 10 п.н.). GC_3 рассчитывался в кодирующих последовательностях генов с помощью пакета seqinr R. Степень наклона GC_3-градиента оценивалась с использованием линейной регрессии между GC_3 и положением относительно первого кодирующего нуклеотида (ATG) на основе первых 1000 п.н. сплайсированной последовательности транскрипта. GC_3-бедные и GC_3-богатые гены выделялись по 10% и 90% квантилям величины GC_3. Стабильность ДНК-дуплекса оценивалась с помощью PromPredict [206] в скользящем окне размером 15 п.н. Все манипуляции с координатами последовательностей производились с помощью пакета bedtools и пользовательских скриптов на языках R и C [204]. Поиск известных сайтов связывания транскрипционных факторов выполнялся с использованием базы данных TRANSFAC и программы MATCH [207].

2.4 Разработка и апробация микросателлитных маркеров

2.4.1 Идентификация повторов и дизайн праймеров

Для поиска тандемных повторов была использована сборка полного генома *L. sibirica* (NCBI GCA_004151065.1; [12]). При помощи GMATo v1.2 [161] был проведен отбор контигов, содержащих микросателлитные повторы с необходимыми параметрами: тандемные повторы с длиной мотива от 3 до 6 п.н. и минимальным числом повторений мотива 15 раз для трехнуклеотидных, 10 для четырехнуклеотидных, 7 раз для пяти- и шестинуклеотидных мотивов. Для всех потенциальных микросателлитных локусов было задано ограничение на расположение не ближе, чем 20 п.н. к началу и концу контига [30]. Дизайн праймеров был выполнен в онлайн-сервисе WebSat (https://bioinfo.inf.ufg.br/websat/). При проведении гель-электрофореза фрагментов, которые различаются всего на несколько нуклеотидов, с большей надежностью определяются те, что имеют относительно небольшую длину. Поэтому при отборе последовательностей праймеров предпочтение отдавалось целевому размеру продукта 140–280 п.н.

Поскольку поиск повторов выполнялся в полногеномной сборке, была проведена проверка на наличие среди выявленных повторов тех, что принадлежат хлоропластному и митохондриальному геномам. Для этого с помощью BLAST было проведено выравнивание контигов, содержащих повторы с подобранными праймерами, на сборку хлоропластного генома и черновую сборку митохондриального генома лиственницы сибирской. Выровнявшиеся контиги затем дополнительно проверялись выравниванием на базу NCBI для уточнения их сходства с митохондриальными последовательностями других растений [30].

Для отбора специфичных праймеров (т.е. таких, которые не имеют повторений в геноме) их последовательности были выравнены при помощи BLAST (task blastn, perc_identity 100) против исходных контигов *L. sibirica*. Все последовательности, имевшие более одного точного совпадения, удалялись. Прошедшие фильтрацию

последовательности праймеров были синтезированы в ЗАО «Евроген», и использовались для дальнейшего тестирования [30].

2.4.2 Отбор полиморфных маркеров

На первом этапе отбирались праймеры с успешной амплификацией, а также проводилась оптимизация условий ПЦР. Биологическим материалом послужила выборка хвои лиственницы сибирской, собранная в 2016 г. близ поселка Туим Республики Хакасия. Способность праймеров давать амплификат проверялась на образцах ДНК четырех деревьев *L. sibirica,* с последующим проведением электрофореза в полиакриламидном геле и окрашиванием продуктов ПЦР при помощи раствора бромистого этидия [30]. В работе использовалась программа ПЦР с «touchdown» (Таблица 3) для уменьшения неспецифической амплификации [30]. После оптимизации условий амплификации каждый локус тестировался на 8–10 образцах из одной выборки лиственницы сибирской с целью выявления полиморфизма и наличия нуль-аллелей. По результатам данной проверки были отобраны локусы, демонстрирующие наиболее стабильные интерпретируемые спектры [30].

Таблица 3 — Программа ПЦР амплификации

Этап амплификации	Температура	Время	Количество циклов
Первичная денатурация	94°C	1 мин	1
Денатурация	94°C	30 сек	9
Отжиг праймеров	60°C, с понижением на 1°C каждый цикл	30 сек,	
Элонгация	72°C	1 мин	
Денатурация	94°C	30 сек	24
Отжиг праймеров	50°C	30 сек	
Элонгация	72°C	30 сек	
Денатурация	72°C	10 мин.	1
Охлаждение	4°C	–	–

Поскольку иногда локусы, оказавшиеся мономорфными для конкретной выборки, могут являться полиморфными для другого вида, изменчивость всех отобранных перспективных локусов проверяли далее на восьми образцах от трех

видов из географически отдаленных популяций — *L. sibirica*, *L. gmelinii* и *L. cajanderi* [30].

Препараты тотальной ДНК были выделены модифицированным методом с применением цетилтриметиламмониумбромида (CTAB) из образцов тканей хвои, высушенной при помощи силикагеля (Рисунок 6) [208]. Для проведения ПЦР использовали готовые реакционные смеси для амплификации ДНК «GenePak PCR Core» производства ООО «Лаборатория Изоген» (Москва, Россия), содержащие ингибированную для «горячего старта» *Taq*-ДНК-полимеразу, дизоксинуклеозидтрифосфаты и хлорид магния [30]. Продукты амплификации разделяли путем электрофореза в 6%-ом полиакриламидном геле с использованием трис-EDTA-боратного электродного буфера в камерах для вертикального фореза. Окраску геля проводили в растворе бромистого этидия с последующей визуализацией в ультрафиолетовом свете. Маркером стандартных длин служила ДНК плазмиды pBR322 *E. coli*, обработанная рестриктазой *Hpa*II (Рисунок 6) [30].

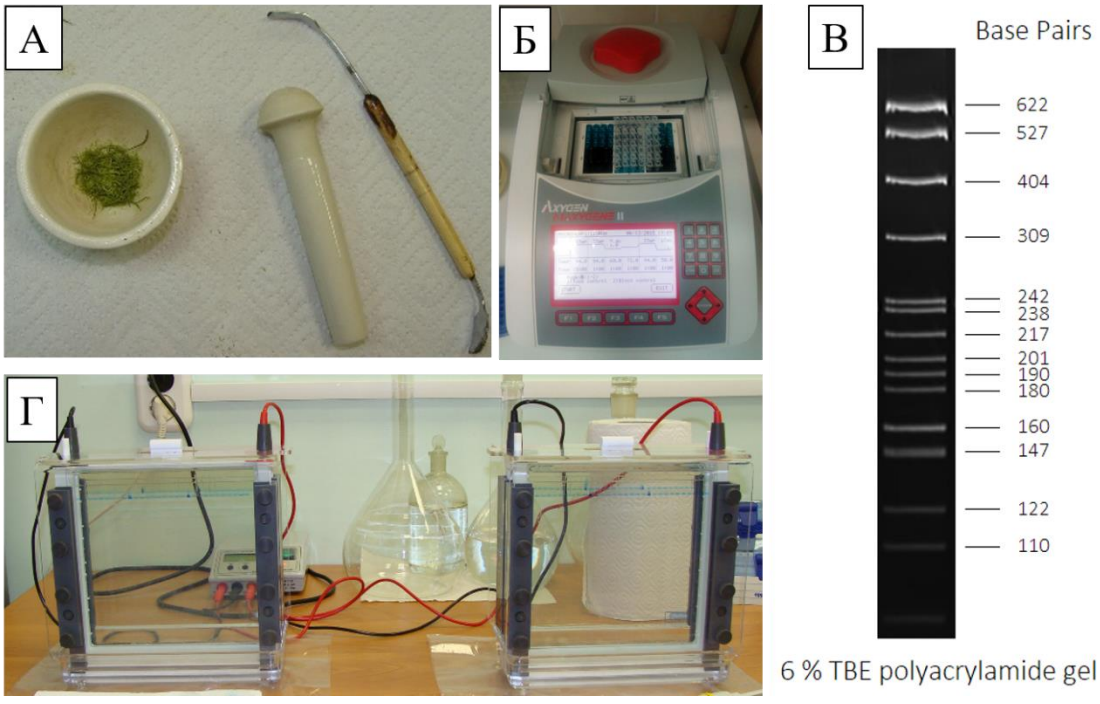


Рисунок 6. Лабораторное оборудование, используемое для проведения тестового популяционно-генетического исследования лиственницы: **А** — высушенная хвоя в ступке для измельчения, **Б** — амплификатор и стрипы с образцами для проведения ПЦР, **В** — маркер стандартных длин *Hpa*II, **Г** — вертикальный электрофорез в полиакриламидном геле (фото автора).

Для окончательного тестирования локусов были взяты выборки из четырёх географически отдаленных популяций трех видов лиственницы, представляющих

разные части их ареалов [30]. Названия популяций и их местоположения представлены в Таблице 4.

Таблица 4 — Географическое расположение четырёх популяций лиственницы, использованных для тестирования отобранных микросателлитных локусов

Вид	Популяция	Код популяции	Район расположения	Географические координаты
L. sibirica	Северо-Енисейск	ЛС-СЕ	Северо-Енисейский район Красноярского края, Пит-Городок.	59º 22′ с.ш. 93º 39′ в. д.
	Туим	ЛС-ТУ	Окрестности села Туим, Республика Хакасия	54º 13′ с.ш. 89º 59′ в.д.
L. cajanderi	Якутия	ЛК-ЯК	Республика Якутия (Саха), Намский район, окрестности с. Хомустах-1 (Кысыл Сыр).	62º 46′ с.ш. 129º 69′ в.д.
L. gmelinii	Чита	ЛГ-Ч	Выборка из природной популяции на территории Забайкальского края, в окрестностях населенного пункта Хилок.	51º 21′ с.ш. 110º 27′ в.д.

Стоит отметить, что в работе мы не заостряем внимание на дискуссионных вопросах систематики видов рода *Larix* и придерживаемся взглядов А.П. Абаимова и И.Ю. Коропачинского, по мнению которых лиственницы Гмелина и Каяндера — два разных вида [209,210]. Объем популяционных выборок всех видов лиственницы составил 24 дерева.

2.4.3 Оценка показателей генетического разнообразия

Для оценки уровня генетического разнообразия были рассчитаны среднее число аллелей на локус (A), средняя наблюдаемая (H_o) и ожидаемая гетерозиготности (H_e) и эффективное число аллелей (N_E). Среднее число аллелей на локус (N_A) оценивалось путем деления общего числа обнаруженных аллелей на число исследуемых локусов [30]. Наблюдаемая гетерозиготность (H_o) рассчитывалась как отношение числа гетерозиготных образцов к общему числу проанализированных по данному локусу образцов. Расчет показателя ожидаемой гетерозиготности производился по формуле:

$$H_e = 1 - \sum f_k^2, \quad (3)$$

где f_k — частоты аллелей, выявленных для k-го локуса.

Эффективное число аллелей было рассчитано по формуле:

$$n_e = \frac{1}{(1-H_e),} \quad (4)$$

где H_e — уровень гетерозиготности по всем проанализированным локусам.

Для оценки популяционной структуры и степени генетической подразделенности популяций были рассчитаны показатели *F*-статистик Райта [211]. Индекс фиксации в субпопуляциях (F_{IS}):

$$F_{IS} = 1 - \frac{H_o}{H_{\acute{e}}} \quad (5)$$

где H_o — наблюдаемая гетерозиготность; $H_{\acute{e}}$ — средняя ожидаемая гетерозиготность. Индекс фиксации особи относительно вида (F_{IT}) вычислялся из соотношения:

$$F_{IT} = F_{IS} + (1 - F_{IS})F_{ST} \quad (6)$$

Коэффициент межпопуляционной дифференциации (F_{ST}) был рассчитан для каждого *i*-того аллеля конкретного локуса, и далее усреднялся для всего локуса:

$$F_{ST} = \frac{H_T - \acute{H}_E}{H_T}, \text{ в которой } H_T = 1 - \sum \acute{p}_i^2 \quad (7)$$

где $\acute{p}_i$— средняя частота *i*-го аллеля. Значения данных параметров, рассчитанные для каждого локуса отдельно, далее усреднялись по всем локусам.

Количественная оценка уровня дивергенции между исследованными популяциями была определена с использованием стандартного генетического расстояния М. Нея [212] по формуле:

$$D = -lnI_N, \text{ в которой } I_N = \frac{\sum\sum_{jk}^{x} Y_{jk}}{\sqrt{\left(\sum\sum x_{jk}^2\right)\left(\sum\sum y_{jk}^2\right)}}, \quad (8)$$

где I — показатель генетического сходства; x_{jk} и y_{jk} — частоты *j*-го аллеля *k*-го локуса в сравниваемых популяциях. Все показатели были рассчитаны при помощи программного обеспечения GenAlEx 6.5 [30,213]. Для проверки правильности генотипирования использовалась программа Micro-Checker [147]. С ее помощью была проведена оценка частот возможных нуль-аллелей в локусах и скорректированы число гомозиготных генотипов и частота амплифицированных аллелей в популяциях. Частоты скрытых нуль-аллелей рассчитывались согласно правилу Харди-Вайнберга.

ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1 Ядерный геном лиственницы сибирской

3.1.1 Анализ высокоповторяющихся элементов генома

К основным факторам, ответственным за увеличение размера генома у высших растений относятся полиплоидия и пролиферация мобильных элементов, которые не только увеличивают размер генома, но также вносят изменения в кодирующие и регуляторные области генов, повышая скорость мутаций и изменяя паттерны экспрессии генов. В литературе обсуждается необратимость процесса накопления повторов в геномах покрытосеменных и хвойных [5,17]. Мнения по этому поводу противоречивы: одни исследователи считают это результатом резкого повышения пролиферации мобильных элементов [214–218], в то время как другие предполагают, что накопление повторов могло происходить постепенно с течением времени, конкурируя при это с процессом элиминации повторов, что могло также сокращать размера генома [5,219,220].

Видоспецифичная библиотека повторов для лиственницы сибирской, созданная *de novo* с помощью RepeatModeler, содержала 1721 консенсусную последовательность. С помощью сборки консенсусных последовательностей скриптом Inchworm из пакета TrinityRnaSeq удалось получить ~31 000 консенсусных последовательностей, которые, с высокой долей вероятности, представляют повторяющиеся участки генома лиственницы сибирской. Эти консенсусные последовательности были сравнены с *de novo* библиотекой RepeatModeler, а также с библиотекой повторов PIER. Гомологи были найдены для ~12 000 консенсусных последовательностей в библиотеке, полученной из RepeatModeler, и для ~7 000 последовательностей в базе данных PIER. Реципрокный BLAST показал, что 1045 из 1721 последовательностей, полученных с помощью RepeatModeler, имели близкую гомологию с консенсусными последовательностями, полученными с помощью кластеризации [20]. Отдельная библиотека RepeatModleler для лиственницы сибирской, а также комбинированная библиотека повторов, депонированы в figshare с DOI 10.6084/m9.figshare.19785913,

и размещены в облачном хранилище вычислительного кластера СФУ по адресу https://hpccloud.sfu-kras.ru/owncloud/index.php/s/PMnkYcj8Lbb1X5R [20].

Относительная представленность классифицированных семейств повторов в геноме лиственницы сибирской аналогична ранее описанному для других хвойных. Общее количество повторяющихся элементов в сборке генома составило 20,9 млн с общим размером 4,8 млрд. п.н, что составляет около 40% от размера генома (Приложение А). Доля генома, покрытая повторами в части длинных ридов Oxford Nanopore, по оценке RepeatMasker, составила 66% п.н. [20].

Использование TEclass позволило лучше реконструировать группы и семейства мобильных элементов, составляющих большую часть генома. Ретротранспозоны класса I надсемейства *LINE, I, Gypsy* и *Copia* оказались наиболее многочисленными, причем элементы *LINE* также имеют самый длинный средний размер и занимают наибольшую часть генома (Приложение А, Рисунок 7А) [20].

Класс I Длинные концевые ретротранспозоны (LTR), представленные в основном элементами *Gypsy* и *Copia*, составляют наибольшую часть всех мобильных элементов. Значительная часть LTR была гомологична библиотекам искусственных бактериальных хромосом (BAC) сосны ладанной и последовательностям фосмид [169,221]. Были идентифицированы *PtTalladega* (3 646 копий в геноме лиственницы сибирской), *PtOuachita* (1025), *IFG* (990), *PtAppalachian* (773), *PtConagree* (731) и еще восемь семейств повторов гомологичных повторам сосны ладанной. Однако большинство LTR-ретротранспозонов не были классифицированы в более мелкие семейства («неклассифицированные LTR» в Приложении 1) [20].

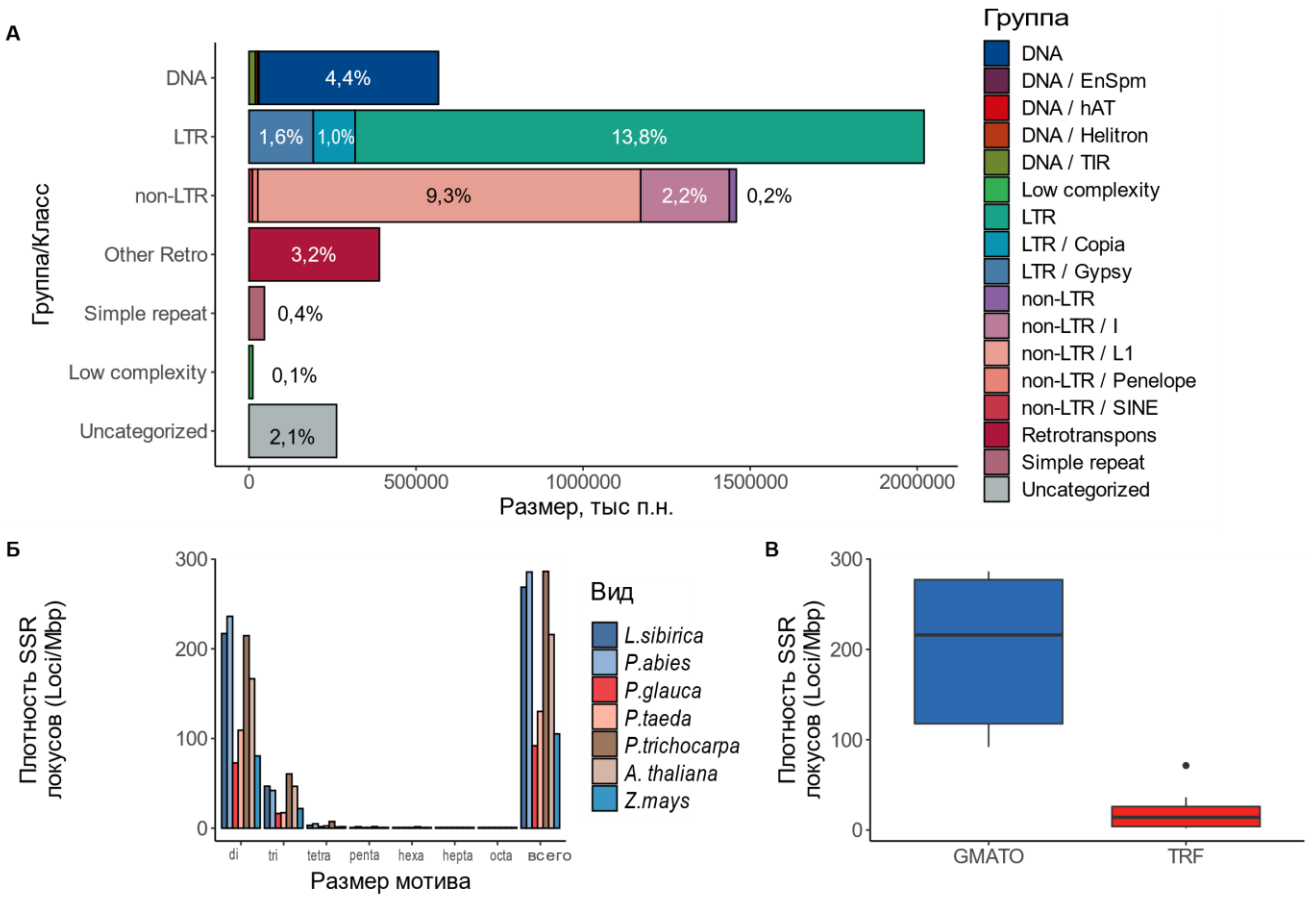


Рисунок 7. Повторы в геноме лиственницы сибирской: **А** — относительное содержание семейств повторов в геноме лиственницы сибирской; **Б** — плотность микросателлитных локусов (число локусов с ди-, три-, тетра-, пета-, гекса-, гепта- и октонуклеотидными мотивами на 1 млн. п.н.) для нескольких видов хвойных покрытосеменных видов, оцененная с помощью GMATo; **В** — среднее число микросателлитных локусов найденное с использованием GMATo и TRF.

Среди non-LTR-ретротранспозонов *LINE/L1*, *I*, *Penelope* и *SINE* вместе составляют около 98% всех non-LTR-ретротранспозонов, которые покрывают 12% длины сборки. Большинство повторов среди разных семейств повторов были относительно небольшими по длине, менее 1 тыс. п.н. Небольшая часть самых длинных повторов достигала почти 15 тыс. п.н.; они принадлежат элементам *LINE* и неклассифицированным LTR. Наиболее частые повторы для каждого семейства были короче 1 тыс. п.н. Некоторые группы повторов имеют бимодальное распределение длины (*Gypsy*, *DIR*, *LINE/I*, *Helitron*, *Penelope*), но оба пика в распределениях были менее 1 тыс. п.н. [20].

LTR составил наибольшую долю всех подвижных элементов с преобладанием семейства *Gypsy*, что также широко наблюдается у других хвойных [14,17,54,55,222] и покрытосеменных [223]. В то время как *LINE* и *SINE* являются

общими для геномов растений, элементы подобные семейству *Penelope* (*PLE*) долгое время считались особенностью геномов животных и грибов [224,225], пока множественные *Penelope (EN(+)* элементы типа *PLE* и *Dryad* не были обнаружены также и у сосны обыкновенной [226], а элементы типа *RTE AdLINE3* — у ряда видов цветковых растений [227]. Описанный в литературе филогенетический анализ семейств *Dryad* и *AdLINE3* предполагает горизонтальный перенос мобильных элементов, возможно, между членистоногими и предком хвойных примерно 340 млн лет назад [226,227].

DNA транспозоны класса II покрывают около 5% размера сборки, большая часть из которых (4,5%) не были классифицированы с помощью TE-class («неклассифицированные» в Приложении 1) [20]. Среди классифицированных транспозонов наиболее многочисленны терминальные инвертированные повторы (*TIR*, 0,16% DNA транспозонов), *Helitron* (0,06%), *EnSpm* (0,02%), *hAT* (<0,01%).

Всего с помощью GMATo в геноме лиственницы сибирской было обнаружено 1 129 244 микросателлитных локуса с размером мотива 2-8 п.н. при средней плотности микросателлитов 268,7 локусов на миллион п.н. По сравнению с другими видами, сборка генома лиственницы также имела относительно высокую плотность микросателлитов, аналогичную геномам ели европейской и тополя черного (Рисунок 7Б) [20]. В работах [55] и [21] сообщалось о плотности SSR 10–20 локусов/Mbp для сосны ладанной, ели белой и ели обыкновенной, при оценке с помощью TRF. На тех же геномных сборках TRF идентифицировал 17 145 локусов с тем же размером мотива и с общей плотностью 4,1 локуса/Mbp. В среднем GMATo обнаружил в девять раз больше локусов SSR, чем TRF, на основе семи видов растений (Рисунок 7**В**; в среднем 197 и 21 локус/Мбп для GMATo и TRF соответственно) и оказался более эффективным для обработки больших геномных последовательностей [20].

Хотя типы повторов и их распределение в геноме лиственницы сибирской соответствуют таковым у других хвойных растений, только 40% ее текущей геномной сборки составляют простые повторы и мобильные элементы. Однако, оценка RepeatMasker показала, что доля генома, покрытого повторами в порции

длинных прочтений Oxford Nanopore, составляет 66% п.н. Это может указывать на то, что репитомная часть генома лиственницы сибирской была слишком фрагментирована для включения ее в окончательную сборку [20]. Эта оценка тем не менее ниже, чем у всех остальных голосеменных растений. Похожие данные были в 2021 году получены для двух других видов лиственницы, *L. decidua* и *L. kaempferi* [228], что может говорить о том, что сравнительно меньшая доля повторов в геноме характерна для видов рода *Larix*.

3.1.2 Оценка времени вставки ретротранспозонов LTR-RT

Ретротранспозоны класса I размножаются за счет интеграции своей промежуточной РНК в геном хозяина посредством ретротранскрипции в кДНК с использованием механизма транскрипции хозяина и собственных ферментов. Между длинными концевыми повторами находится кодирующая часть, содержащая гены *gag* и *pol*. [229,230]. Эта кодирующая часть включает протеазу, обратную транскриптазу, рибонуклеазу-H и интегразу, которые отвечают за расщепление белка Pol и РНК, копирование РНК ретротранспозонов в кДНК и интеграцию кДНК в геном хозяина (Рисунок 8А) [231].

У высших растений преобладают повторы с прямыми LTR [171,232]. Фланкирующие LTR на 5’ и 3’-концах повторов идентичны в момент вставки, но со временем они мутируют [233], при этом частота их мутаций, по всей видимости, выше, чем в кодирующих областях, поскольку повторы в отличие от генов не находятся под давлением отбора. Количество различий между двумя фланкирующими LTR можно использовать для оценки времени вставки элемента в геном. Это может помочь понять эволюционные аспекты организации генома и обнаружить недавние и древние события повторной экспансии.

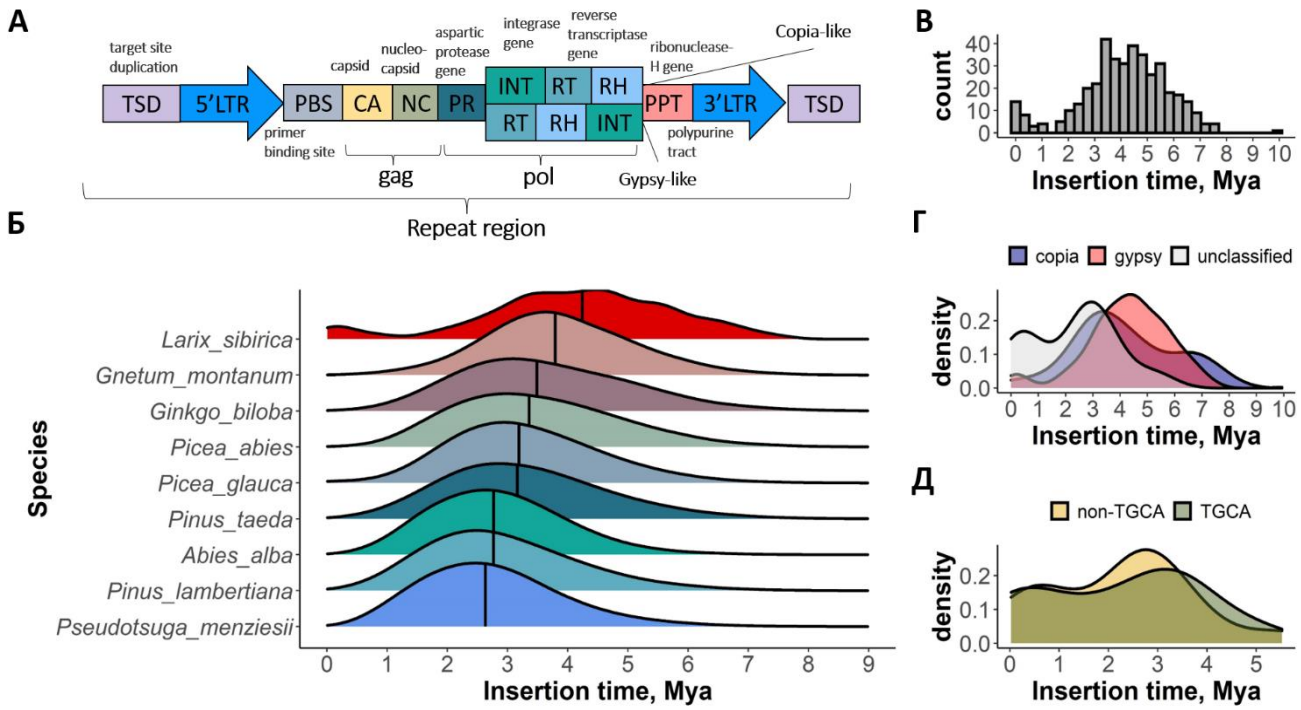


Рисунок 8. LTR повторы: **А** — структура Copy- и Gypsy-подобных LTR ретротранспозонов; **Б** — оценка времени вставки LTR-RT элементов в геноме девяти покрытосеменных; **В** — оценка времени вставки LTR-RT элементов в геноме лиственницы сибирской; **Г** — оценка времени вставки суперсемейств Copia и Gypsy; **Д** — оценка времени вставки TGCA/non-TGCA LTR элементов. По оси абсцисс — время вставки в млн. лет.

LTRharvest с последующей обработкой LTR_retriever идентифицировали 347 LTR элементов и 36 интактных LTR в сборке лиственницы сибирской. Эти 36 интактных LTR были объединены с 367, идентифицированными [173]. Возможное перекрытие было проверено с помощью blastn против LTR *Larix* из [173]. Оценка количества замен в концевых повторах 403 LTR элементов позволяет предположить, что вероятная волна встраивания ретротранспозонов в геном лиственницы произошла порядка 4–5 млн лет назад (Рисунок 8В) [20]. Хотя суперсемейства Copia (PR-INT-RT) и Gypsy (PR-RT-INT) имеют немного разные профили оценки времени встраивания, их средние и медианные значения близки (среднее = 3,16 млн лет назад и медиана = 3,03 млн лет назад для Copia, среднее значение = 3,11 млн лет назад и медиана = 2,96 млн лет назад для Gypsy) (Рисунок 8Г). LTR с разными фланкирующими мотивами, типичными 5'-TG…CA-3', и другими менее распространенными вариантами мы также сравнивали с точки зрения времени их вставки. Точно так же LTR с фланкирующими мотивами TG…CA имели небольшой пик с медианой 2,56 млн лет назад, а LTR с другими

фланкирующими мотивами имели медиану 2,60 млн лет назад (Рисунок 8Д). По сравнению с профилями других голосеменных, лиственница показала самый древний всплеск LTR, даже по сравнению с гнетумом и гингко (Рисунок 8Б) [20].

Содержание остаточной ДНК LTR было намного выше, чем содержание интактных LTR, что позволяет предположить, что после массивной пролиферации ретротранспозонов в геноме лиственницы сибирской могла произойти элиминация повторяющейся ДНК. Типичные оценки времени вставки в геномы растений варьируются от 1 до 2,5 млн лет назад для покрытосеменных [234–238]. Сообщается, что у голосеменных растений время внедрения составляет 10–15 млн лет назад [51]. На основании идентификации LTR, проведенной [173] приблизительное время экспансии LTR голосеменных можно оценить в 2–4 млн лет назад. У лиственницы на оценку времени может влиять либо эффективный механизм элиминации повторов в сочетании с истинной вставкой древних повторов, либо фрагментарный характер черновой сборки и, следовательно, малое количество найденных LTR. Однако черновые геномы ели европейской и пихты белой имеют сравнимую степень цельности сборки (N50 = 6 443, 5 206 и 14 051 п.н. для лиственницы сибирской, ели европейской и пихты, соответственно), и, несмотря на заметное различие в числе идентифицированных LTR (403 у лиственницы, 31 016 у ели европейской и 34 952 у пихты), оценки времени встраивания LTR для них также сходны.

3.1.3 Идентификация генов с лейцин богатыми повторами (LRR)

Многие белки содержат богатые лейцином повторы (LRR), которые имеют подковообразную структуру и могут участвовать в белок-белковых взаимодействиях. Особенно выделяются белки NBS-LRR, которые играют важную роль в защите растений от патогенов. Участки LRR имеют высокую изменчивость, что обеспечивает специфичность распознавания молекул патогенов, и вероятно связаны с иммунным ответом растений на биотический стресс [239–242].

Среди всех тканях было обнаружено 4 482 транскрипта, содержащих LRR-домен. Наибольшее количество LRR-доменов содержалось в транскриптоме побега (1 846), несколько меньше в транскриптоме камбия (1 599), но их доли в общем

количестве транскриптов для каждой ткани были примерно одинаковыми для всех тканей и составляли не более 2%. Как видно на рисунке в Приложении 2, наибольшее количество транскриптов содержало семейство LRR-4 во всех тканях [20].

Белки NBS-LRR имеют длину от 860 до 1 900 аминокислот, но большинство транскриптов, содержащих домен LRR, были короче на 300–400 аминокислот (Приложение Б). Все последовательности короче 850 были отфильтрованы и среди них был проведен поиск домена NBS. Всего в транскриптоме побега было обнаружено 56 предполагаемых белков NBS-LRR, 18 в камбии, 5 в проростке и 2 в осенней почке. OmicsBox подтвердил наличие доменов NB-ARC и LRR в транскриптах, а InterProScan не обнаружил других функциональных доменов в этих последовательностях. Вероятно, эти последовательности содержат гены устойчивости к заболеваниям и стрессу, так как они включают семейства P-loop NTPase и LRR [20]. Последовательности, содержащие идентифицированные домены LRR и NB-ARC, размещены в figshare (DOI 10.6084/m9.figshare.19785913) и в облачном хранилище суперкомпьютерного кластера СФУ https://hpccloud.sfu-kras.ru/owncloud/index.php/s/PMnkYcj8Lbb1X5R.

3.1.4 Структурная аннотация с использованием программы MAKER2

Работа с такими большими геномами, как у хвойных, часто затруднена из-за ограниченности вычислительных ресурсов, таких как время вычислений и объем памяти, необходимый для обработки геномных данных. Структурная аннотация полногеномной сборки лиственницы сибирской с помощью пайплайна MAKER2 на кластере из 448 ядер (с тактовой частотой 2,3 ГГц/ядро и 896 ГБ оперативной памяти при средней загрузке процессора около 61%) заняла 22 дня, исключая настройку AUGUSTUS и подбор базы данных повторов.

Сравнительный анализ с помощью BUSCO выявил 317 полных и 307 фрагментированных генов из 1 614 однокопийных ортологов. Относительно высокая доля фрагментированных генов (19% фрагментированных генов против 38,6% всего найденных) может объясняться большой долей фрагментированных скаффолдов в сборке. При этом в геномных сборках других хвойных относительное

количество фрагментированных генов по оценке BUSCO в среднем ниже (7,5% против 80,9% для *Pinus lambertiana*, 11,5% против 32,6% для *Picea glauca*, Таблица 5) [20].

Таблица 5. Оценка полноты генного состава для нескольких видов группы хвойных с помощью BUSCO в режиме сравнивания белковых последовательностей.

Вид	Полные, %	Частичные, %	Полные и частичные, %
Pinus lambertiana	73,4	7,5	80,9
Pseudotsuga menziesii	68,5	11,8	80,3
Pinus taeda	41,7	19,4	61,1
Picea abies	28,1	27,3	55,4
Larix sibirica	19,6	19,0	38,6
Abies alba	15,8	17,9	33,7
Picea glauca	21,1	11,5	32,6

Используя транскрипты из нескольких тканей, транскриптомные сборки родственных видов хвойных и референсные белки Uniprot в качестве отправной точки для аннотаций MAKER2, было получено 39 370 моделей генов в 37 206 скаффолдах, состоящих из 134 271 экзона и 94 901 интрона (Таблица 6). Среди них 24 551 генная модель была полноразмерной, а 14 819 — частичными (6 476 усеченных с начала, 7 545 усеченных с конца и 798 усеченных с обеих сторон). Средняя длина генов составила около 1 841 п.н., один ген содержал в среднем 3,41 экзона, при этом наиболее частым числом экзонов было 2, что согласуется с прогнозом ~4 экзона на ген для *Pinus taeda* [243]. Максимальная длина CDS составила 7 216 п.н., что меньше длины самого длинного интрона, 10 153 п.н. (Таблица 6) [20].

Таблица 6. Статистика геномной сборки и аннотации лиственницы сибирской.

Параметр	*Larix sibirica*
Количество хромосом	12
Размер генома, Gbp [244]	12,03
Размер сборки, Gbp	5,59 [а] / 12,34 [б]
N50, bp	3 098 [а]/ 6 443 [б]
GC состав, %	35,41
Содержание повторов, %	66
Количество предсказанных генных моделей	39 370
Количество полно-длинных генных моделей	24 551
Средняя длина CDS, bp	244,29
Средняя длина интрона, bp	360,93
Максимальны длина интрона, bp	10 153

Примечание: сборка в [а] контигах, [б] скаффолдах

Количество предсказанных моделей генов у лиственницы сибирской (39 370) примерно такое же, как у сосны обыкновенной (50 172) и пихты Дугласа (54 830), но значительно меньше, чем у пихты белой (94 205), сосны сахарной (71 117), ели белой (102 915) и ели европейской (70 968) [20]. MAKER2 использует показатель контроля качества, называемый редакционным расстоянием аннотации (AED), который впервые был представлен в проекте Sequence Ontology [245,246]. Он измеряет соответствие между моделью гена и сопутствующими подтверждающими данными, вместо оценки расстояния между аннотациями, как это было сделано в проекте Sequence Ontology [177]. По оценке MAKER2 AED, рассчитанное для аннотации лиственницы сибирской, было ниже 0,5 для 95% моделей генов, что сравнимо с таковым для генома мыши GRCm37 [177]. Однако, учитывая малое количество подтверждающих данных для данного вида, которые можно использовать в предсказании генов и для контроля качества, эта оценка может быть в некоторой мере завышена.

Для областей, идентифицированных RepeatMasker как повторы, также были обнаружены пересечения с кодирующими участками из предсказанных генных моделей. Всего 6 884 гена имели по крайней мере 20% перекрытия с повтором (Приложение B). Эти генные модели были помечены как «связанные с повторами»; 2 247 (33%) из них пересекались с семейством Non-LTR I, 241 (3%) с LINE, 571 (8%) с LTR Gypsy, 523 (8%) с Copia и 312 (5%) с простыми повторами. Наиболее частыми функциональными аннотациями для генов с перекрывающимися повторами были рецептороподобные протеинкиназы, белки, богатые лейцином, повторы (LRR), факторы транскрипции, АТФ-связывающие переносчики, ферменты синтазы, редуктазы, эстеразы и пероксидазы, ферменты цитохрома С, белки цитохрома Р450 и другие (Приложение B) [20].

Так же, как и размеры генома, средняя длина интрона также больше у хвойных, чем у покрытосеменных растений [247]. В аннотации, полученной с помощью MAKER2, 94 901 интрон был идентифицирован в общей сложности в 36 183 генах со средней длиной 361 п.н. и самым длинным интроном 10 153 п.н., что меньше, чем у других видов хвойных; 289 интронов были длиннее 5 тыс. п.н.

[20]. При сравнении 10% самых длинных интронов, интроны лиственницы были сопоставимы по длине с таковыми у *A. thaliana* и *P. taeda*, хотя самые длинные интроны лиственницы были намного короче, чем у других видов ели *P. abies* и *P. glauca*, или в богатых повторами геномах *Populus thicocarpa, Vitis vinifera* и *Zea mays* (Рисунок 9).

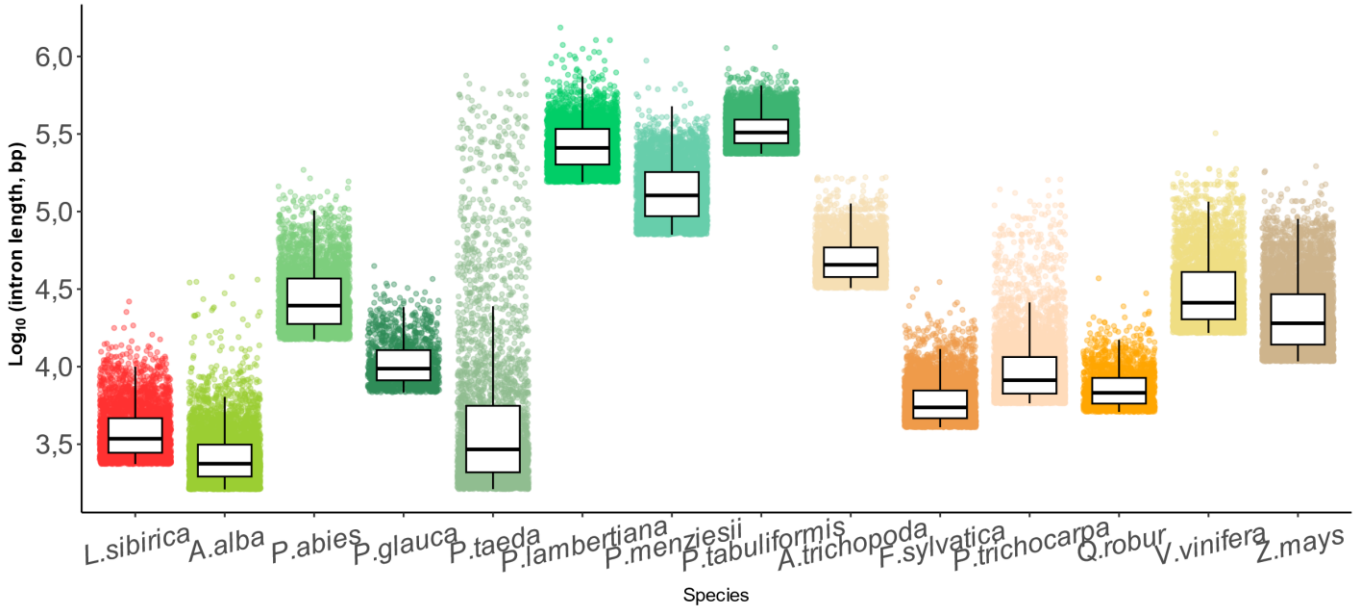


Рисунок 9. 10% самых длинных интронов в 11 видах покрытосеменных и хвойных растениях

В сумме интроны составляли до 47 % (34,25 млн. п.н.) генного пространства и 0,29 % сборки генома. Содержание повторов в интронах было ниже, чем в геноме в целом, только 4,59 млн. п.н. (13% интронного пространства) покрыты повторами. В интронах лиственницы наиболее распространёнными оказались ретротранспозоны класса I LINE и I (6 135 и 2 195 элементов соответственно), за которыми следуют LTR Gypsy и Copia (1 879 и 1 214 элементов соответственно). Среди ДНК-транспозонов класса II наиболее часто встречались элементы TIR и EnSpm (362 и 245 элементов соответственно) [20].

Средняя длина интрона в геноме лиственницы была в 1,8–3,2 раза короче, чем у других хвойных. При сравнении верхних 10% самых длинных интронов, интроны лиственницы были намного короче, чем у других видов хвойных, таких как *P. abies, P. glauca* и *P. taeda* (Рисунок 9). Это расхождение может быть объяснено (1) недооценкой размера интрона пайплайном MAKER2 [16], который использует встроенные пороговые значения для разрешения экзон-интронной структуры, или

(2) естественными различиями внутри группы хвойных. Цельность сборок у *L. sibirica* и *P. abies* близка, о чем свидетельствует их N50 (Таблица 7), а их средняя и максимальная длина интрона различаются в 2,8 и 6,7 раза соответственно. Точно так же средний размер интрона у *L. sibirica* и *P. glauca* был близким, различаясь в 1,8 раза (Таблица 7), а их N50 различались в 7,2 раза. Тем не менее, возможно, что различия в размерах интронов в геноме хвойных могут быть объяснены разницей в цельности сборок [20].

Таблица 7. Статистика геномных сборок и генных аннотаций для некоторых видов хвойных и покрытосеменных

Параметр	*L. sibirica*	*P. taeda*	*P. abies*	*P. glauca*	*P. trichocarpa*[а]
Размер генома, Gbp	12,03	20,15	19,57	15,79	0,48
Размер сборки, Gbp[б]	12,34	22,10	12,30	25,47	0,42
N50, bp[б]	6,443	110,557	7,747	46,559	–
Количество хромосом	12	12	12	12	19
GC состав (%)	35,41	38,06	38,81	37,08	
Содержание повторов, %	65,98	81,8	70,0[в]	–	41
Количество предсказанных генных моделей	39,370	50,172	70,968	102,915	41,377
Количество полно-длинных генных моделей	24,551	–	–	–	–
Средняя длина CDS, bp	244,29	419,81	287,21	283,56	233,05
Средняя длина интрона, bp	360,93	1146,12	997,94	642,73	468,08
Максимальны длина интрона, bp	10,153	568 968	68 268	44 113	96 842

Примечание: [а] сборка хромосомного уровня, [б] скаффолды, [в] на основе несобранных прочтений [17].

3.1.5 Функциональная аннотация

87% предсказанных моделей генов лиственницы (34 358 из 39 370) имели гомологию с белками *Arabidopsis thaliana* при минимальных пороговых значениях e-value ≤ 10^{-5}, покрытие ≥ 20% и идентичность ≥ 20%) (Рисунок 10). Доля картированных белков у лиственницы была выше, чем у большинства других голосеменных растений, но ниже, чем у сосны китайской *P. tabuliformis* и некоторых модельных покрытосеменных растений, таких как тополь черный, виноград культурный, дуб обыкновенный и бук европейский (Рисунок 10Б). При этом, при обратном картировании 72% белков Арабидопсиса (19 706 из 27 416, при минимальных пороговых значениях покрытия ≥ 20% и идентичности ≥ 20) имели гомологов среди белков лиственницы [20].

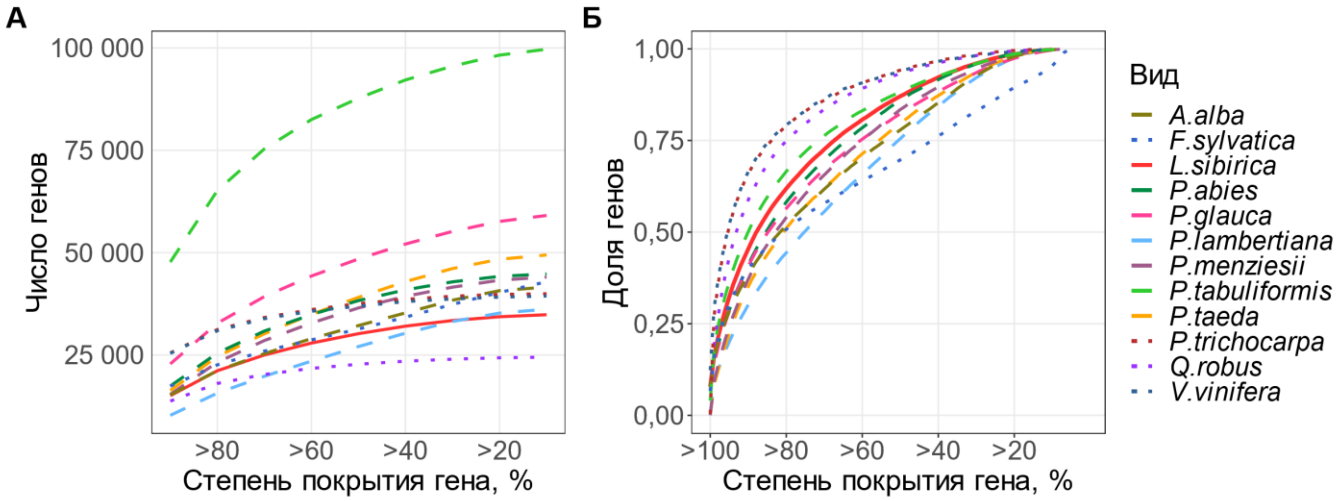


Рисунок 10. Гены, имеющие сходство с белками арабидопсиса (покрытие qcovhsp выше указанного порога, минимальная идентичность 20%): **А** — кумулятивное количество, **Б —** доля в процентах. Голосеменные отмечены штрих-линией, покрытосеменные — пунктиром, лиственница сибирская — сплошной линией.

Присвоение категории GO было основано на идентификации доменов с помощью InterProScan и поиске гомологии с BLAST. В результате было получено 30 512 аннотированных моделей генов (78%), по крайней мере, с одним присвоенным термином GO. Для более детального анализа все гены были разделены на 20 функциональных категорий. Функции классифицированы в соответствии с последней версией словаря генной онтологии: 5 категорий в биологическом процессе, 6 в молекулярной функции, 5 в клеточном компоненте (Рисунок 11А). Все белки из соответствующей категории были картированы на базу белков *Arabidopsis* с помощью blastp ($e \leq 10^{-5}$, pident> 50 и qcovhsp> 50). От 50% (в категории транскрипционной активности) до 85% (в категориях транспортная активность, митохондрия и хлоропласт) аннотированных белков лиственницы сибирской оказались гомологами белков арабидопсиса (Рисунок 11Б) [20].

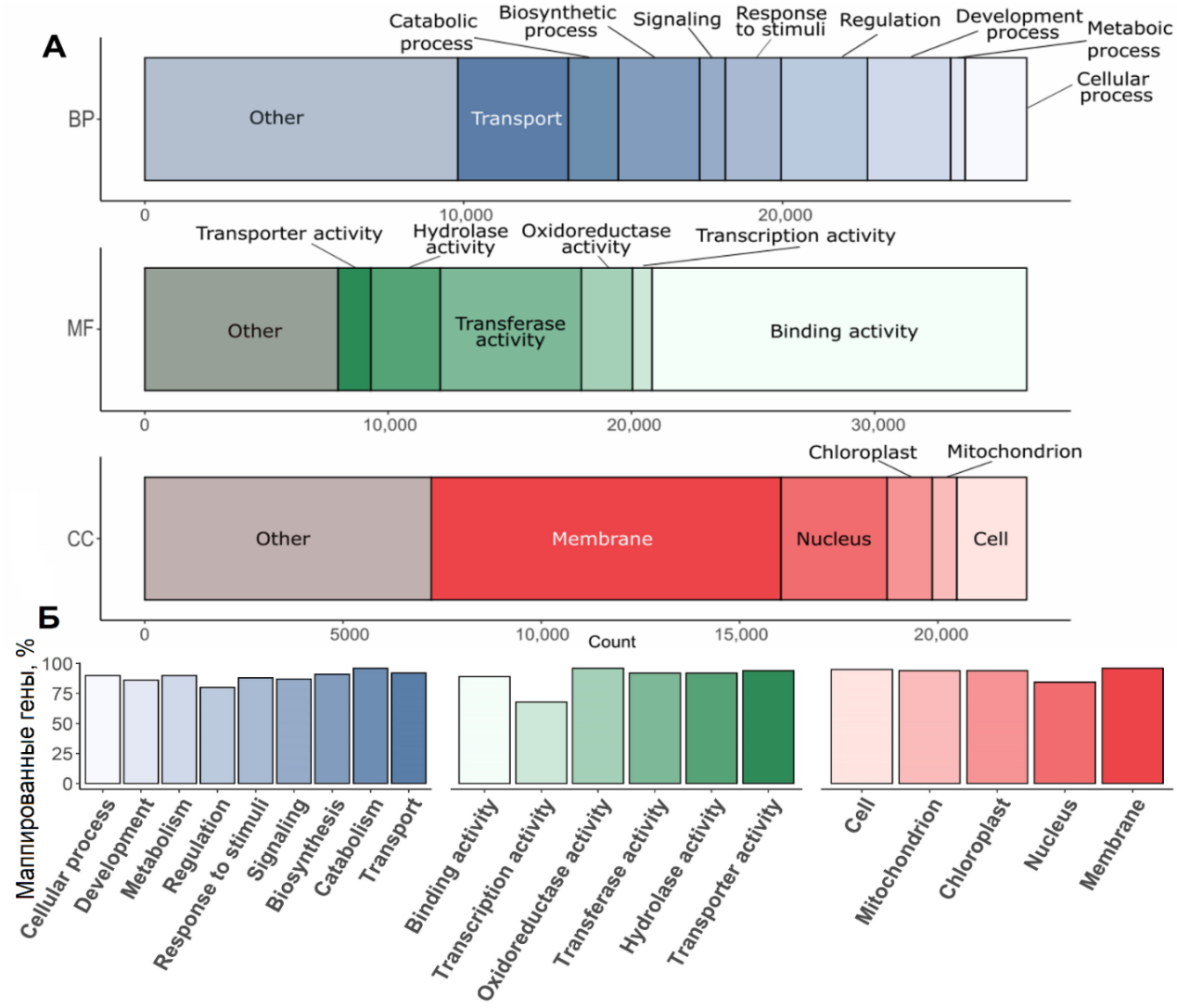


Рисунок 11. Функциональная аннотация генов лиственницы сибирской**: А —** доля предсказанных генов в трех функциональных категориях (BP — биологический процесс, МФ — молекулярная функция и CC — клеточный компонент); **Б —** процентное содержание белков лиственницы в различных функциональных категориях, картированных на белки арабидопсиса (BLASTP e ≤ 10^{-5}, pident > 20 and qcovhsp > 20).

Хвойные отличаются от цветковых древесных растений рядом признаков, в том числе отсутствием сосудов в ксилеме и ситовидных трубок и клеток-спутников во флоэме, иной структурой древесины, гаплоидным мегагаметофитом (в отличие от триплоидного эндосперма у покрытосеменных), мегаспорофиллами, репродуктивными структурами, представленными шишками, а не цветками. Хвойные, в отличие от покрытосеменных, в основном вечнозеленые растения, и очень немногие из них имеют сезонное опадение иголок (сезонное старение), которое встречается у видов родов *Glyptostrobus*, *Metasequoia*, *Taxodium*, *Pseudolarix* и *Larix*.

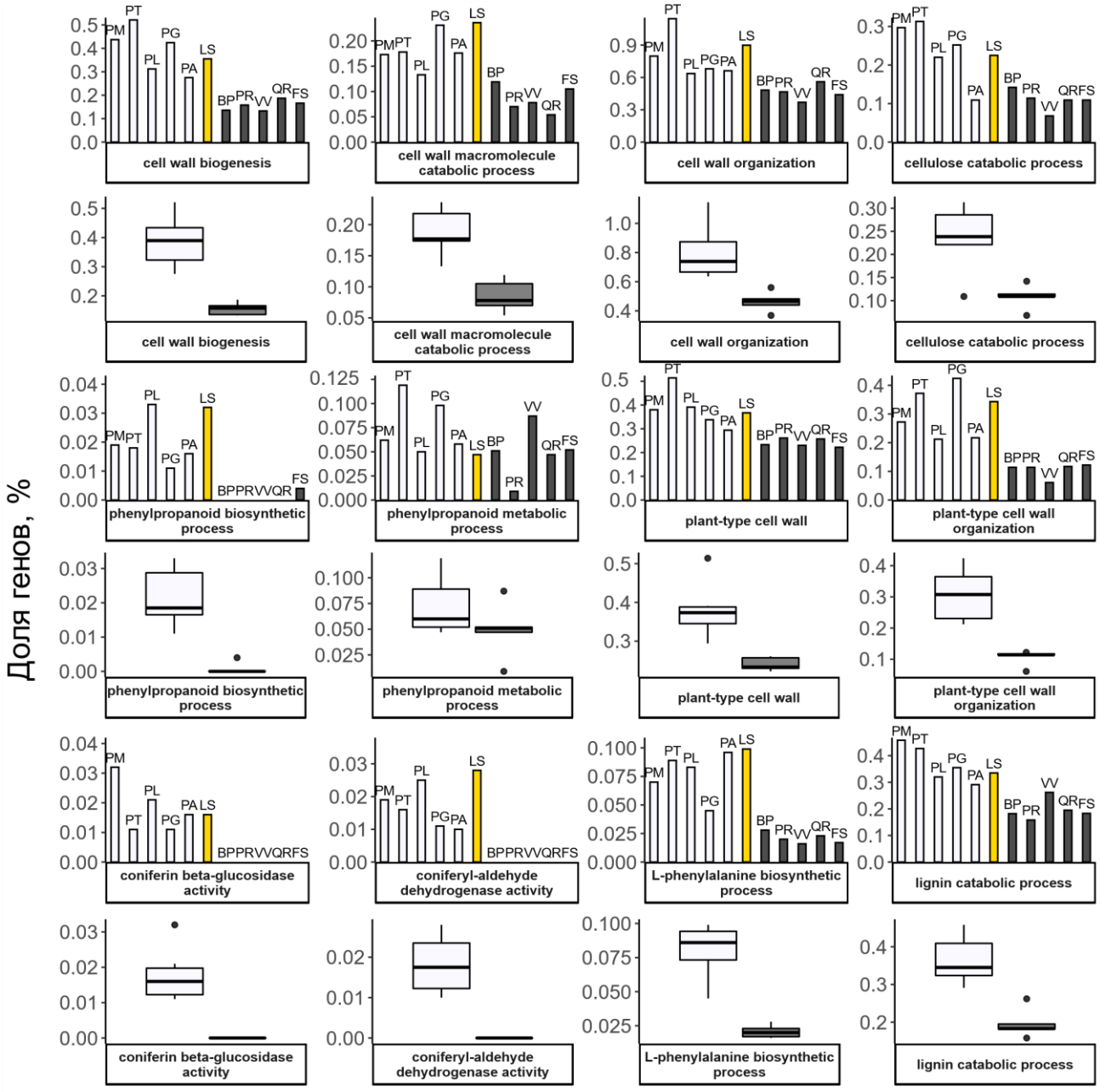
3.1.5.1. Клеточная стенка и метаболизм фенилаланина

Структура клеточных стенок оказывает влияние на свойства древесины, которая содержит гемицеллюлозу, пектин (в первичной клеточной стенке) и лигнин (во вторичной клеточной стенке) [248]. У покрытосеменных древесных растений состав лигнина вторичной клеточной стенки отличается от хвойных, где основными строительными блоками являются р-кумариловый спирт и конифериловый спирт. [249–252].

Различия между хвойными и покрытосеменными древесными видами можно четко увидеть по количеству генов в терминах GO, связанных с организацией клеточной стенки и катаболизмом лигнина (Рисунок 12). Ферменты клеточной стенки, участвующие в биосинтезе лигнина из р-кумарилового и конифорилового спиртов, идентифицированы у всех шести проанализированных видов хвойных [20].

Одним из основных предшественников биосинтеза лигнина и производства вторичных метаболитов является фенилаланин. Хвойные используют больше углерода для биосинтеза лигнина и фенилпропаноидов, которые играют важную роль в защите от насекомых и микробных патогенов [253–255].

Биосинтез фенилаланина и фенилпропаноидов также демонстрирует различие между хвойными и покрытосеменными растениями **(**Рисунок 12) [20]. У растений основными ферментами, участвующими в биосинтезе фенилаланина, являются префенат-аминотрансфераза (PAT), которая превращает префенат в арогенат, и арогенатдегидратаза (ADT), которая превращает арогенат в фенилаланин, который, наконец, превращается фенилаланин-аммиак-лиазой (PAL) в коричную кислоту в цитозоле, первый компонент фенилпропаноидного пути [253,256]. Ранее было показано, что хвойные имеют более разнообразные семейства генов ADT и PAT по сравнению с покрытосеменными [257,258].



Покрытосеменные: BP — *Betula pendula*; FS — *Fagus sylvatica*; PR — *Populus trichocarpa*; QR — *Quercus robur*; VV — *Vitis vinifera*. Голосеменные: PM — *Pseudotsuga menziesii*; PT — *Pinus taeda*; PL — *Pinus lambertiana*; PG — *Picea glauca*; PA — *Picea abies*; LS — *Larix sibirica*.

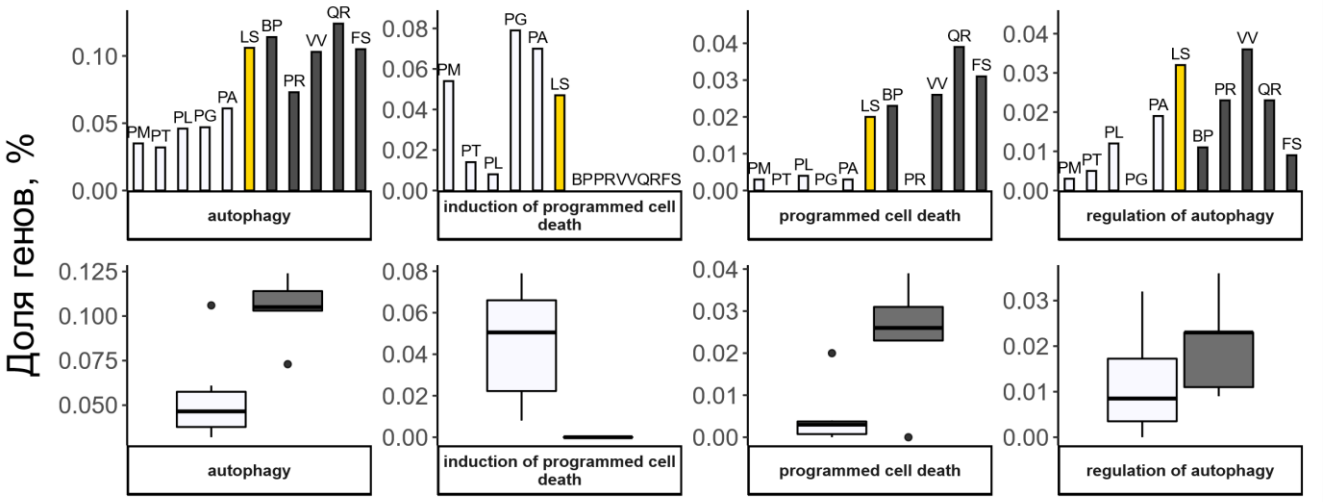
Рисунок 12. Процент генов, аннотированных терминами GO, связанными с развитием и строением клеточной стенки. Листопадные покрытосеменные представлены черными серыми столбцами, вечнозеленые голосеменные — белыми, лиственница сибирская — желтым. Диаграммы размаха демонстрируют разницу в количестве генов между двумя группами: вечнозелеными (белый) и листопадными (серый).

3.1.5.2. Запрограммированная гибель клеток и аутофагия

Программируемая клеточная гибель (PCD), также известная как апоптоз у животных, является процессом регулируемого клеточного самоубийства, который может быть вызван внешними факторами или происходить по физиологическим

причинам в процессе развития организма. Растительные клетки имеют жесткую клеточную стенку, которая препятствует образованию апоптотических телец и не обладают классическими каспазами, действующими как основные индукторы PCD, или фагоцитозом [259]. Вместо этого они используют каспазоподобные протеазы (метакаспазы) для индукции PCD [260] и вакуоли для переваривания содержимого своих клеток [261,262]. PCD у растений можно разделить на два основных типа. Первый тип — вакуолярная, также известная как аутолитическая или апоптозоподобная PCD, характеризующаяся поглощением цитоплазмы литическими вакуолями и последующим высвобождением вакуолярных гидролаз в цитозоль вследствие разрыва тонопласта. Второй тип — PCD гиперчувствительного ответа, которая характеризуется сморщиванием клеток и повышенной аутофагической активностью [259,263–265]. Первый тип обычно происходит во время дифференцировки элементов ксилемы, старения листьев и мегаспорогенеза, а второй активируется в ответ на инвазию патогена для предотвращения дальнейшего распространения инфекции [265]. Было показано, что PCD и аутофагия тесно связаны со старением [266,267].

В категории GO:0012501, связанной с PCD, лиственница сибирская имеет число генов, более близкое к таковому у листопадных покрытосеменных деревьев, чем у вечнозеленых хвойных. Количество генов, связанных с аутофагией, также выше у листопадных покрытосеменных и лиственницы, чем у других хвойных (Рисунок 13). Однако гены группы GO:0012502 (индукция PCD) были аннотированы только для хвойных видов (20 у дугласовой пихты, 6 у сосны обыкновенной, 2 у сахарной сосны, 42 у ели белой, 22 у ели европейской и 12 у сибирской) [20].



Покрытосеменные: BP — *Betula pendula*; FS — *Fagus sylvatica*; PR — *Populus trichocarpa*; QR — *Quercus robur*; VV — *Vitis vinifera*. Голосеменные: PM — *Pseudotsuga menziesii*; PT — *Pinus taeda*; PL — *Pinus lambertiana*; PG — *Picea glauca*; PA — *Picea abies*; LS — *Larix sibirica*.

Рисунок 13. Процент генов, аннотированных терминами GO, связанными с запрограммированной гибелью клеток (PCD) и аутофагией. Листопадные покрытосеменные представлены черными серыми столбцами, вечнозеленые голосеменные — белыми, лиственница сибирская — желтым. Диаграммы размаха демонстрируют разницу в количестве генов между двумя группами: вечнозелеными (белый) и листопадными (серый).

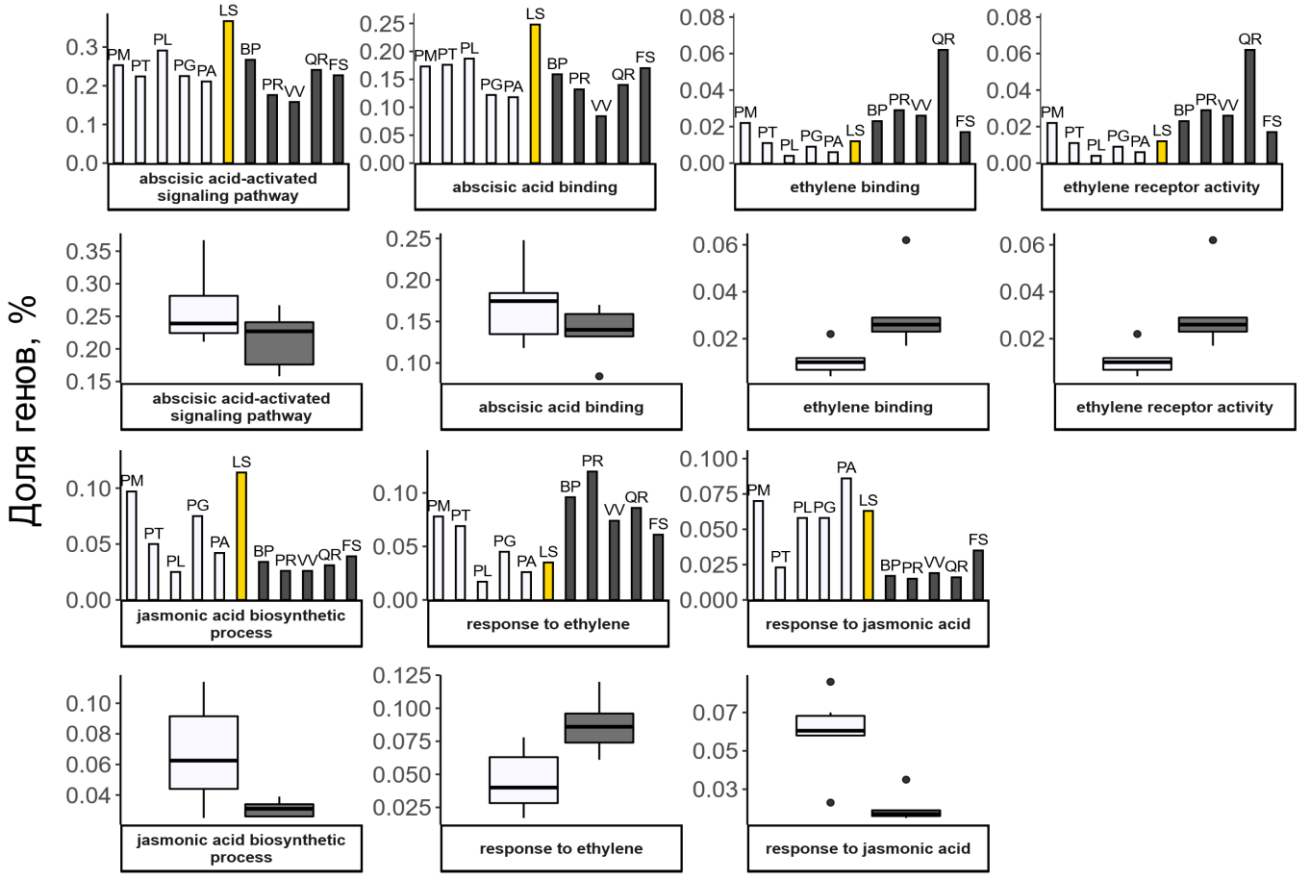
Известно, что обработка этиленом запускает PCD в клетках зоны опадания, а применение поглотителей активных форм кислорода замедляет опадание [268]. Было высказано предположение, что PCD регулируется через сигнальный путь этилена, поскольку мутанты с дефектом биосинтеза этилена демонстрируют повышенную продолжительность жизни листьев [269]. Активация аутофагии в процессе PCD и старения была показана на мутантах арабидопсиса с дефицитом аутофагии, которые демонстрировали ускоренное старение и PCD [270,271].

3.1.5.3. Гормоны

Абсцизовая кислота (АБК), жасмоновая кислота (ЖАК) и этилен являются важными фитогормонами, играющими важную роль в ответе на абиотический и биотический стресс и старении листьев. АБК важна для устойчивости к абиотическим стрессам и заражению патогенами; она вызывает закрытие устьиц, тем самым уменьшая потерю воды через транспирацию в ответ на дефицит воды или тепловой стресс [272]. Мутанты арабидопсиса с дефицитом АБК имеют пониженную устойчивость к холоду [273–276]. На примере многих растений было показано, что АБК участвует в старении листьев [277,278]. Обработка риса и

арабидопсиса экзогенной АБК ускоряет пожелтение и старение листьев [279,280], а уровни эндогенной АБК увеличиваются во время старения листьев у кукурузы и арабидопсиса [281,282].

Лиственница сибирская имеет наибольшее количество аннотированных генов, связанных с реакцией на гормоны, процессом биосинтеза ЖАК, сигнальным путем, активируемым АБК, и связыванием АБК (Рисунок 14). Напротив, у всех хвойных, включая лиственницу сибирскую, количество генов связывания этилена, активность рецепторов этилена и реакция на этилен относительно ниже, чем у большинства покрытосеменных [20]. Это наблюдение можно назвать ожидаемым, учитывая, что последние компоненты канонического сигнального пути этилена появились уже после отделения покрытосеменных от голосеменных [283,284].



Покрытосеменные: BP — *Betula pendula*; FS — *Fagus sylvatica*; PR — *Populus trichocarpa*; QR — *Quercus robur*; VV — *Vitis vinifera*. Голосеменные: PM — *Pseudotsuga menziesii*; PT — *Pinus taeda*; PL — *Pinus lambertiana*; PG — *Picea glauca*; PA — *Picea abies*; LS — *Larix sibirica*.

Рисунок 14. Процент генов, аннотированных терминами GO, связанными с передачей гормональных сигналов и реакцией. Листопадные покрытосеменные представлены черными серыми столбцами, вечнозеленые голосеменные — белыми, лиственница сибирская — желтым. Диаграммы размаха демонстрируют разницу в количестве генов между двумя группами: вечнозелеными (белый) и листопадными (серый).

ЖАК и ее производные также участвуют в регуляции экспрессии генов в ответ на различные абиотические стрессы. Процесс запускается абиотическим стрессом, вызывающим накопление ЖАК в цитоплазме стрессированных листьев [285–287] и активирующим ЖАК-сигнальные пути. Высокий уровень ЖАК активирует гены, чувствительные к жасмонату, которые в норме подавляются комплексом репрессии транскрипции [287]. ЖАК смягчает последствия водного дефицита и засоления почвы [288–290], низкой температуры [237,291], чрезмерного воздействия ультрафиолета [292–294] и участвует в механизмах защиты от патогенов у голосеменных [295,296].

3.2 Органельные геномы лиственницы сибирской

3.2.1 Хлоропластный геном

Общая длина окончательной сборки хлоропластного генома составила 122 560 п.н., что очень близко к 122 474 п.н. у близкородственной европейской лиственницы (*L. decidua*). Полученная сборка была депонирована в GenBank NCBI (NC_036811.1). Аннотация с помощью RAST и путем сравнения с имеющимися аннотациями для *L. decidua* и *L. occidentalis* позволила идентифицировать 110 генов, из которых 34 представляют собой гены тРНК, 4 — рРНК и 72 — белок-кодирующие гены (Рисунок 15) [32].

Геном хлоропластов у хвойных, в том числе у видов лиственниц [297], имеет уникальное строго отцовское наследование через пыльцу, в отличие от покрытосеменных, у которых он имеет материнское наследование через семена [298]. Он позволяет проследить отцовский генный поток и линии отдельно от материнских (митохондриальные гены) и двуродительских (ядерные гены). Поэтому хлоропластная ДНК является крайне важным источником генетических маркеров для изучения распределения отцовских генов и основанных на отцовстве филогенетических отношений у хвойных. Изменчивость генома хлоропластов также можно эффективно использовать для различения разных популяций одного и того же вида. Например, Дж. Б. Уиттол с соавторами [299] продемонстрировали сильную дифференциацию между материковой и островной популяциями сосны

Торри (*Pinus torreyana*) на основе 5 SNPs, обнаруженных во всем хлоропластном геноме размером 120 т.п.н. Однако при средней длине 120–160 т.п.н. и наличии около 130 генов хлоропластные геномы достаточно велики и сложны, включают структурные и точечные мутации, отражающие популяционную дифференциацию и эволюционную дивергенцию [300].

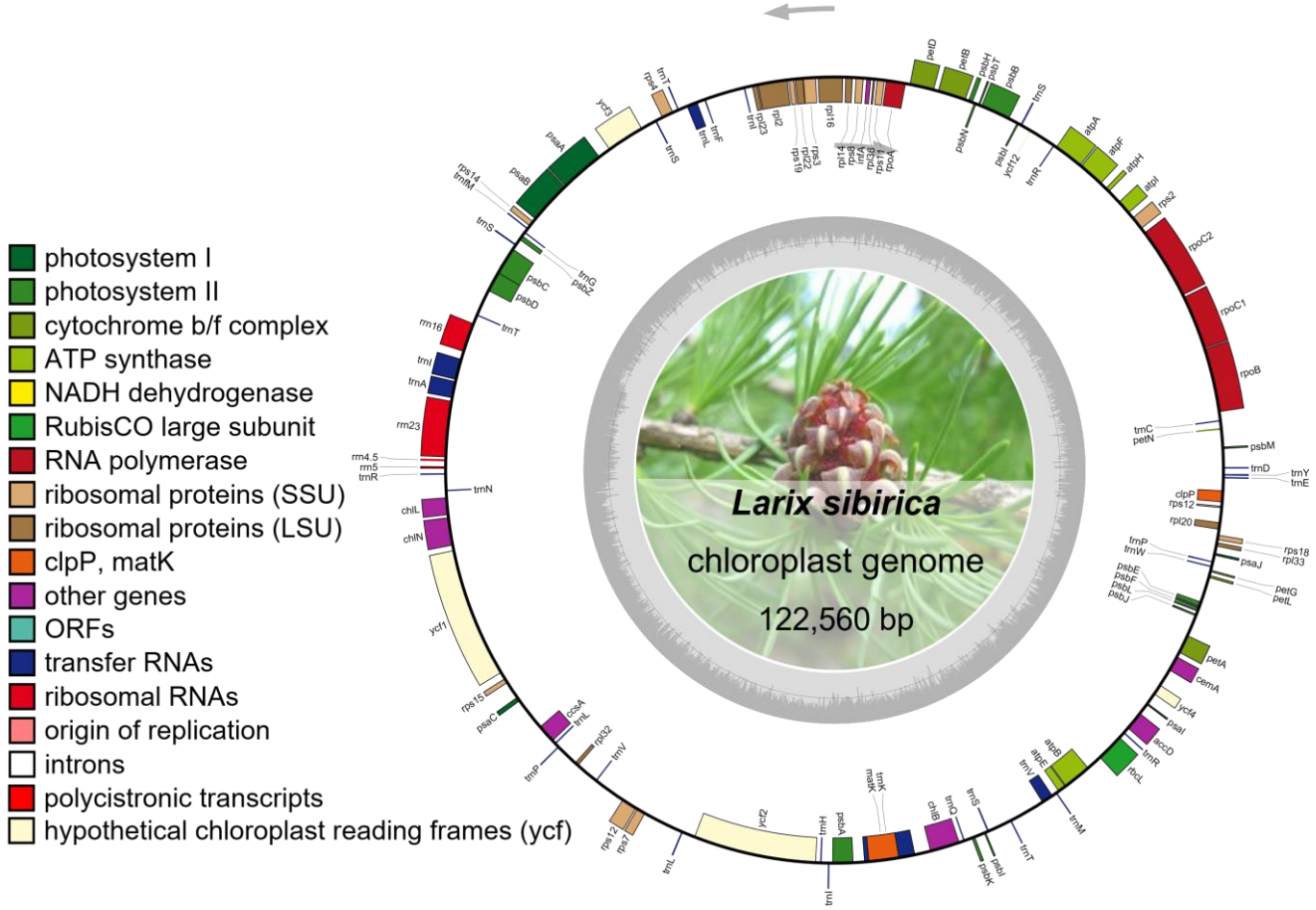


Рисунок 15. Карта расположения генов в хлоропластном геноме лиственницы сибирской. Гены, принадлежащие к разным функциональным группам, отмечены соответствующими цветами. Темно-серым и светло-серым цветами на внутреннем круге изображены GC- и AT-состав, соответственно.

Изменчивость генома хлоропластов у большинства растений часто ограничена из-за относительно низкой частоты мутаций в этой органелле. Например, скорость мутаций хлоропластного генома у сосен составляет примерно $0{,}2–0{,}4 \times 10^{-9}$ синонимичных замен на нуклеотид в год [301,302].

В отличие от покрытосеменных, хлоропластная ДНК (хпДНК) хвойных не имеет больших инвертированных повторов (IR), но содержит диспергированные повторы, связанные со структурными перестройками. Помимо крупных диспергированных повторов, хпДНК хвойных также обладает рядом мелких

повторов — вариабельное число тандемных повторов размером от 124 до 150 п.н., которые связаны с областью полиморфной перестройки вблизи *trnK-psbA*, где имеется дупликация гена *psbA* [303].

Большая часть изменчивости генома хлоропластов связана с микросателлитными локусами [304,305]. Однако эти маркеры имеют слишком высокую частоту мутаций, что может привести к неверным выводам о филогенетических связях [306–308]. SNP могут быть более подходящими маркерами для реконструкции филогенетических связей, и для обнаружения этих маркеров необходимы сравнительные полные исследования геномов хлоропластов.

3.2.2 Митохондриальный геном

На основании ~19,7 миллионов парно-концевых ридов (PE), полученных с помощью секвенатора Illumina HiSeq 2000 из образцов, обогащенных митохондриальной ДНК (мтДНК), и ~625,5 миллионов mate-pair ридов (MP), полученных для полного генома [12], с помощью CLC Assembly Cell и BESST была получена предварительная черновая сборка, которая после картирования на митогеномы семенных растений и фильтрации ядерных и пластидных контигов, составила 53 скаффолда общей длиной 11,09 млн.п.н. Для улучшения сборки митогенома были использованы те же PE риды Illumina, что и в предварительной сборке CLC, а также длинные риды Oxford Nanopore. После сопоставления гибридной сборки MaSuRCa с базой митохондриальных последовательностей растений было собрано девять митохондриальных контигов общей длиной 11,7 млн.п.н. (максимальная длина контига 4 008 762 п.н.). На текущий момент это самый длинный митохондриальный геном из известных [33]. Для дополнительной оценки правильности окончательной сборки был использован REAPR v1.0.18, по результатам которого доля безошибочных нуклеотидов составляла 92,13% от сборки, что сопоставимо с 86% для референсного генома человека GRCh37 или 90,3% для генома *Caenorhabditis elegans* [198]. В ходе аннотации всего было предсказано и аннотировано 40 белок-кодирующих гена, 3 гена рРНК и 31 ген тРНК (Рисунок 16, Таблица 8). Для гена atp8 были обнаружены две копии [33].

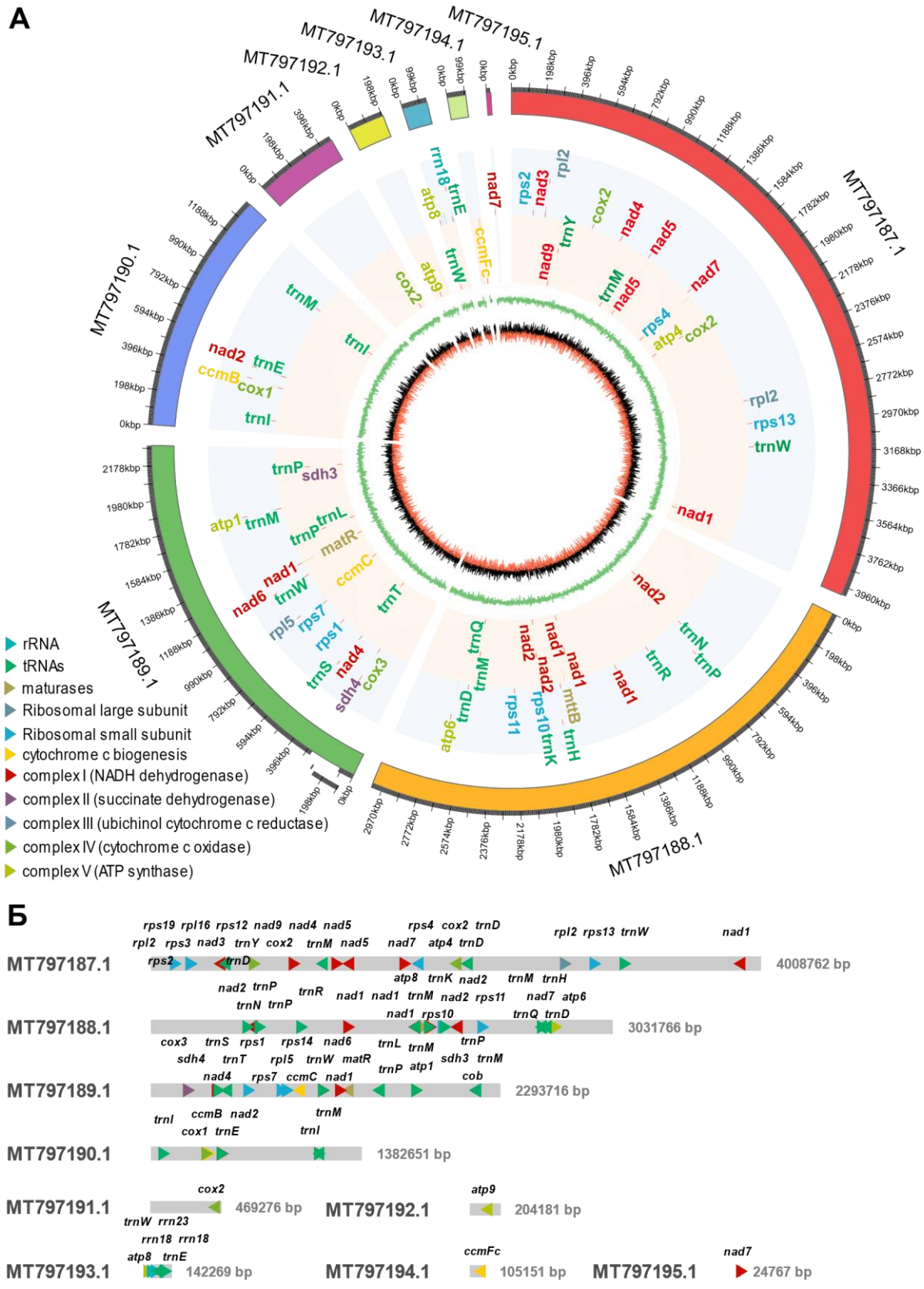


Рисунок 16. Карта расположения генов в митохондриальном геноме лиственницы сибирской. 9 скаффолдов составляют 11,7 млн.п.н. **А** — Круговое отображение. Внешний круг: разным цветом показаны 9 скаффолдов. Внутренние круги: на синем фоне показаны гены прямого стрэнда, на оранжевом — гены обратного стрэнда, зеленым — гистограмма содержания GC, черным/ красным — skewed GC. **Б** — Линейное отображение, серым цветом показаны 9 скаффолдов.

Таблица 8 — Гены, идентифицированные в митохондриальном геноме лиственницы сибирской.

Функциональная группа	Гены
Комплекс I NADH дегидрогеназы	*nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad7, nad9*
Комплекс II сукцинатдегидрогеназы	*sdh3 u sdh4*
Комплекс III убихинол-цитохром с-оксидоредуктазы	*cob*
Комплекс IV цитохром-с-оксидазы	*cox1, cox2, cox3*
Комплекс V АТФ-синтазы	*atp1, atp4, atp6, atp8, atp9*
Гены cytochrome c biogenesis	*ccmB, ccmC, ccmFc, ccmFn*
Гены сплайсинга и транспорта	*matR, mttB*
Гены большой субъединицы рибосомы	*rpl2, rpl5, rpl16*
Гены малой субъединицы рибосомы	*rps1, rps2, rps3, rps4, rps7, rps10, rps11, rps12, rps13, rps194, rps194*

Для обнаружения и классификации часто повторяющихся элементов в собранных скаффолдах были использованы RepeatModeler и TEclass. RepeatModeler обнаружил 122 мобильных элемента, среди которых TEclass идентифицировал 27 ДНК-транспозонов, 38 длинных концевых повторов (LTR), 12 длинных диспергированных ядерных элементов (LINE) и 9 коротких диспергированных ядерных элементов (SINE). Всего с использованием комбинированной библиотеки повторов был идентифицирован 7 691 повтор, что составляет 11,14% от 11,7 млн.п.н сборки. LTR, LINE и простые повторы оказались наиболее распространены среди классифицированных повторов (4%, 2% и 0,8% соответственно; Рисунок 17) [33]. Диспергированные повторы суммарно составляют 7,9% от размера митогенома.

LTR представлены в основном семействами Gypsy, Copia, DIRS и Gymny. Представитель последнего ранее был обнаружен в геноме сосны и родственен элементам Athila у арабидопсиса [309]. Среди non-LTR ретротранспозонов SINE, I и Penelope вместе составляют около 20,56% non-LTR, что составляет 0,53% всего митогенома. RepeatMasker обнаружил три элемента Penelope, все с относительно им score и длиной от 57 до 813 п.н. [33]. Penelope-подобные элементы считаются обычными для животных, но недавно были обнаружены и у хвойных, в частности в геноме сосны обыкновенной [226]. Предполагается, что эти элементы были переданы предку хвойных около 340 млн. лет назад.

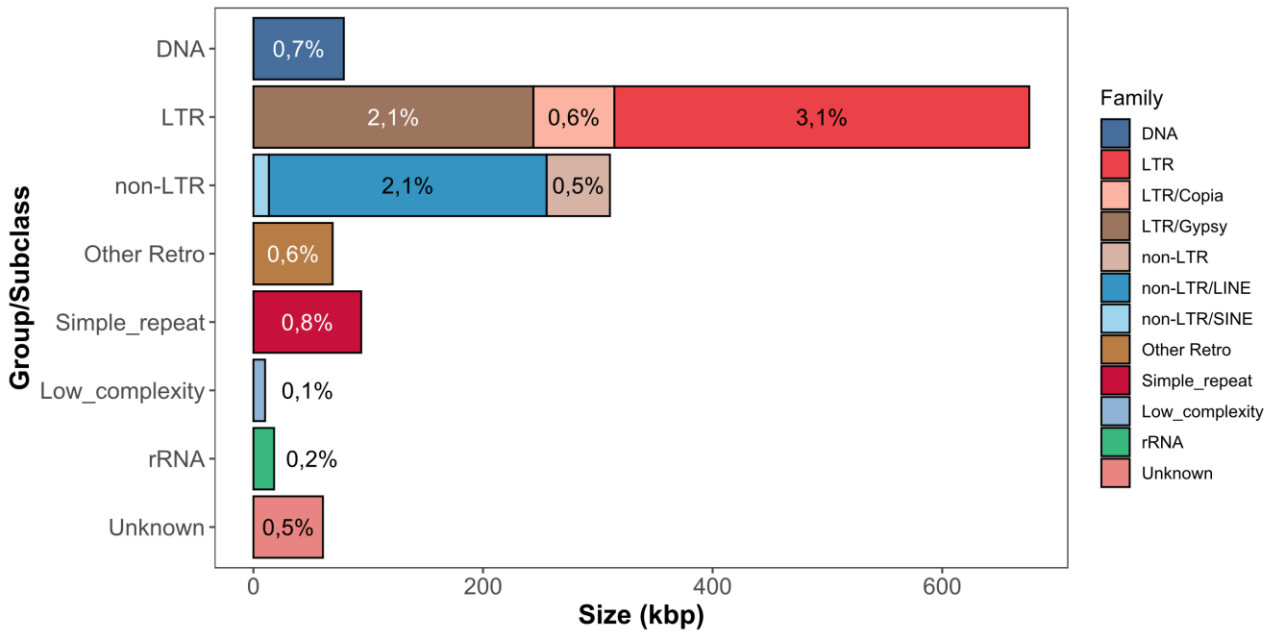


Рисунок 17. Относительное содержание повторяющихся элементов, идентифицированных в митогеноме лиственницы сибирской.

ДНК-транспозоны класса II составляют 0,7% сборки митогенома и представлены семействами повторов EnSpm, Harbinger, hAT, Helitron, MuDR, TcMar и Maverick (Polinton). В сборке митогенома был обнаружен только один элемент Polinton длиной 18 п.н. [33]. Хотя полинтоны не характерны для геномов растений, некоторые элементы, родственные Maverick, обнаружены в цитоплазме и митохондриях некоторых растений и грибов [310,311].

Размер митогенома у семенных растений варьирует по крайней мере на один порядок в диапазоне от ~ 222 т.п.н. у *Brassica napus* [312] и ~ 316 т.п.н. у *Allium cepa* [313] до ~ 3,9 млн.п.н.у *Amborella trichopoda* [314] и ~ 11,3 млн.п.н.у *Silene conica* [315]. Митогеномы растений отличаются от митогеномов животных более высокой сложностью и размером из-за наличия интронов, межгенных последовательностей, повторяющихся и мобильных элементов [316, 317]. Причины чрезвычайно большого размера митогеномов у растений до сих пор полностью не выяснены, но могут быть, по крайней мере, частично объяснены переменным числом мобильных генетических элементов, интронов и связанных с плазмидами последовательностей [320], и могут зависеть от различных факторов, таких как как пролиферация ретротранспозонов, генерация повторяющейся ДНК путем рекомбинации и перенос чужеродных последовательностей из пластидных или ядерных геномов

или посредством горизонтального обмена мтДНК [33,321]. Кодирующие последовательности основных митохондриальных генов относительно консервативны [318,319].

Данная сборка митохондриального генома не состоит из одной последовательности, которую можно было назвать мастер-кольцом, но во многих недавних публикациях показано, что представление митогенома растений в виде единой кольцевой молекулы не всегда точно из-за его сложности и динамического смешения форм мтДНК внутри одного растения [33,322].

Митогеномы как растений, так и животных часто рассматриваются как одна кольцевая последовательность, и до недавнего времени эта модель преобладала при сборке митогеномов. Однако все чаще высказываются предположения о том, что митохондриальные геномы обладают более нетривиальной структурой. Например, митогеном огурца (*Cucumis sativus)* был собран только в три кольцевые хромосомы размером 1,6, 84 и 45 т.п.н. соответственно [323, 324], а митогеном *Silene noctiflora* (6,7 млн.п.н.) фрагментирован более чем на 50 кольцевых хромосом [315]. Интересно, что в митогеномах многих сосудистых растений обнаруживаются чужеродные последовательности, полученные либо путем внутриклеточного переноса генов из пластид [199,325], либо путем горизонтального переноса генов из митохондрий других видов [326]. Большинство этих чужеродных последовательностей являются некодирующими или псевдогенами, хотя иногда встречаются в кодирующих областях [325]. Напротив, перенос гена из митогенома в пластиду считается редким. Более того, принято считать, что гены митохондрий растений эволюционируют медленнее, чем пластидные или ядерные, а частота мутаций в кодирующих областях митогеномов растений примерно на два порядка ниже, чем в генах митохондрий животных [327–330].

Важно отметить, что даже при наилучшем возможном протоколе обогащения мтДНК и наличии достаточного количества интактных молекул мтДНК все равно будет очень сложно проверить, представляет ли конкретная большая сборка митогенома растения единую кольцевую молекулу или нет. Все субгеномные кольца теоретически могут быть обнаружены, если будут получены очень длинные

прочтения, сравнимые с размерами этих субгеномных колец, что пока представляется затруднительным, учитывая огромную длину митогенома хвойных. Для получения однозначных альтернативных сборок длина прочтений должна составлять несколько млн.п.н. Однако даже если бы митогеном был собран как единая кольцевая структура, это не гарантировало бы его действительной кольцевой природы. В митохондриях могут существовать альтернативные физические структуры, способные образовывать некольцевые структуры, такие как конкатамеры «голова к хвосту» и линейные молекулы со специфическими перестановками [33,331].

3.3 Предсказание сайтов начала транскрипции (TSS) в полногеномных сборках ели белой, ели норвежской, сосны ладанной, сосны сахарной и лиственницы сибирской

3.3.1 Предсказание TSS

Для предсказания сайтов начала транскрипции (TSS) были взяты публично доступные аннотации для геномов *Pinus taeda, Picea glauca, Picea abies* и полученная в данной работе аннотация *Larix sibirica*. Для предсказания промоторов был выбран алгоритм TSSPlant, который использует нейронные сети для оценки до 17 признаков, таких как наличие классических мотивов, вариации нуклеотидного состава и другие [205]. Использование TSSPlant позволило предсказать 22 633 позиции TSS у *P. taeda,* 25 889 у *P. abies*, 44 651 у *P. glauca* и 62 420 у *L. sibirica*. От 13,3% до 14,3% идентифицированных позиций TSS находились в кодирующих частях соответствующих моделей генов и были исключены [204].

Чтобы выбрать наиболее вероятную позицию TSS среди нескольких, предсказанных для данного гена, мы сравнили длину каждой 5'-UTR с распределением длин 5'-UTR для четырех видов растений, двух двудольных, *A. thaliana* и *P. trichocarpa*, и двух однодольных растения *O. sativa* и *S. bicolor* (Приложение Г.1). Два параметра *k* и *theta*, определяющие форму и масштаб гамма-распределения, рассчитывались следующим образом: *theta = v/m, k = m/theta*. Используя *k* = 0,62 и *theta* = 238,99, был выбраны предсказания, которые лучше

соответствуют теоретическому распределению длин 5'UTR. После фильтрации предсказаний на предмет попадания в кодирующую область и выбора позиций с наибольшей вероятностью соответствия теоретическому распределению, 10 367 позиций для *P. abies*, 16 629 для *P. glauca*, 9 149 для *P. taeda* и 23 016 для *L. sibirica* были идентифицированы как предполагаемые TSS [204]. Все модели генов с соответствующими предсказанными TSS доступны в геномном браузере Persephone (https://web.persephonesoft.com/).

Аннотации геномов лиственницы сибирской и ели белой были выполнены с помощью пайплайна MAKER с использованием данных транскриптома и доступных данных EST и RNA-seq для этих и других близких видов. Это позволило автоматически предсказать 13 228 UTR для лиственницы сибирской и 14 056 UTR для ели белой на основе доступных данных EST. При сравнении TSS, предсказанных MAKER и алгоритмом TSSPlant с фильтрацией, основанной на распределении длин 5'-UTR модельных видов растений, было показано, что позиционное распределение TSS, предсказанное методом *de novo*, похоже на распределение TSS, предсказанное методом с использованием поддержки РНК (Приложение Г.2) [204].

В предсказанных промоторах появление мотива TATA(A/T)A(A/T) демонстрирует ярко выраженный пик примерно на 20 п.н. выше предсказанного положения TSS для всех четырех видов (Рисунок 18), что хорошо соответствует каноническому расположению TATA-бокса, поскольку эукариотические промоторы содержат TATA-бокс в положении на 40-15 п.н. выше TSS [204].

При сравнении количества промоторов, содержащих мотив TATA-бокс или мотив CA, примерно половина проанализированных последовательностей (46–53%) имела мотив CA в пределах области [–2; +2] вокруг TSS, а TATA-бокс обнаружен у 5–8% промоторов в области [–40; –20] относительно TSS. Среди TATA-содержащих промоторов примерно половина из них содержала мотивы как TATA, так и CA. Растения полагаются на TATA-бокс, чтобы инициировать транскрипцию большинства генов. TATA-бокс расположен примерно на 40 п.н. выше TSS у хвойных. Хотя наиболее распространенное расположение TATA-бокса

находится на расстоянии от 20 до 40 п.н. выше TSS, ранее сообщалось, что у некоторых растений, таких как *Vitis vinifera*, TATA-бокс наблюдался в пределах –70 п.н. относительно TSS [123]. Высота пика частоты TATA напрямую измеряет точность предсказания TSS [204].

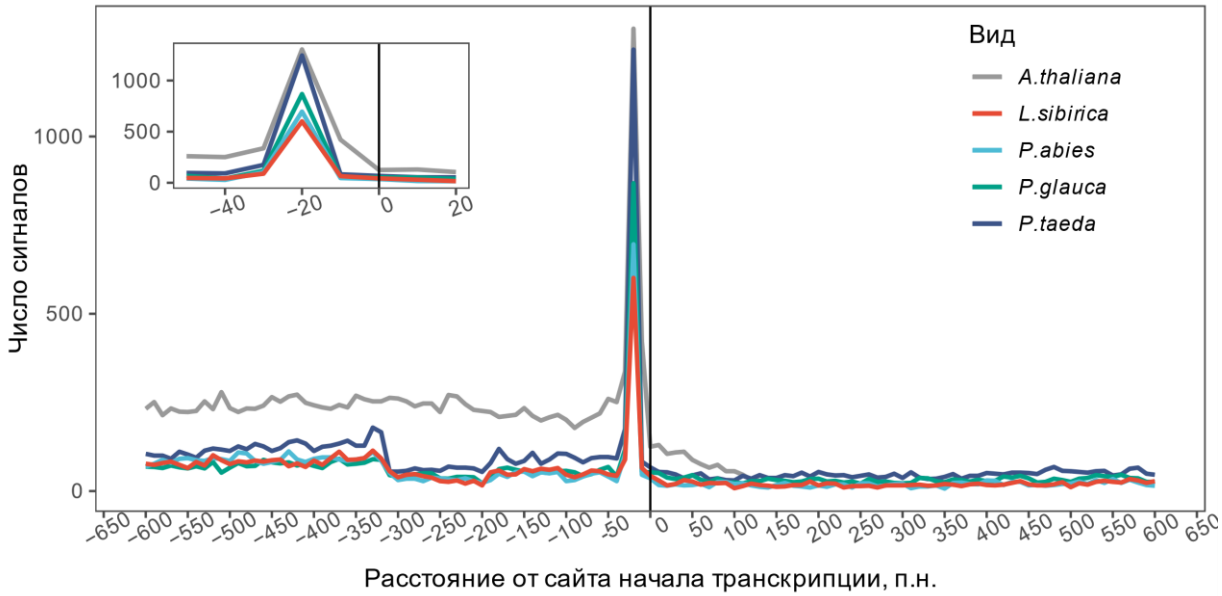


Рисунок 18. Частота мотива TATA(A/T)A(A/T) в TSS-центрированных промоторах для четырех видах хвойных и арабидопсиса.

В дополнение к TATA-боксу стандартная модель базового промотора у растений [332] включает наличие инициаторного мотива в области TSS, AGGA-бокс (YA2-5KNGA2-4YY, ~ 80 п.н. выше TSS, [332,333]), и нижний промоторный элемент, DPE (RGWYVT, ~ 28–32 п.н. ниже TSS, [332]). Существуют принципиальные различия в организации промотора растений и животных. Промоторы млекопитающих обычно имеют CAAT-бокс [334] вместо растительного AGGA. TATA-бокс появляется менее чем в 10% промоторов млекопитающих. Мотив DPE обычно встречается в промоторах без TATA; однако TATA и DPE могут возникать одновременно [332]. Промоторы животных также могут содержать элементы BRE и motif ten (MTE), отсутствующие в промоторах растений [334]. Сообщалось, что гены могут иметь двунаправленные промоторы [335,336]; эта особенность была тщательно изучена у млекопитающих и менее широко у растений. Мортон и соавторы [119] проанализировали образцы корней *A. thaliana*, чтобы найти точное местоположение TSS, и обнаружили, что большинство промоторов не содержат TATA-бокса, но содержат специфическую

комбинацию сайтов связывания факторов транскрипции, регулирующих экспрессию генов. Ямамото и соавторы [337], изучавшие базовые промоторы в геноме *A. thaliana*, такие как TATA и GA, заметили, что архитектура промотора связана со структурой гена. Длина 5'-UTR также отрицательно коррелирует с уровнем экспрессии соответствующих генов [137].

Изменение стандартной свободной энергии ДНК-дуплекса в последовательности генома является индикатором промоторной области и успешно применяется для предсказания промотора. Мы использовали это в качестве подтверждающего доказательства для промоторов, предсказанных TSSPlant. Профиль свободной энергии показывает пик около –40 п.н.. и резкое снижение около предполагаемого TSS (Рисунок 19) [204].

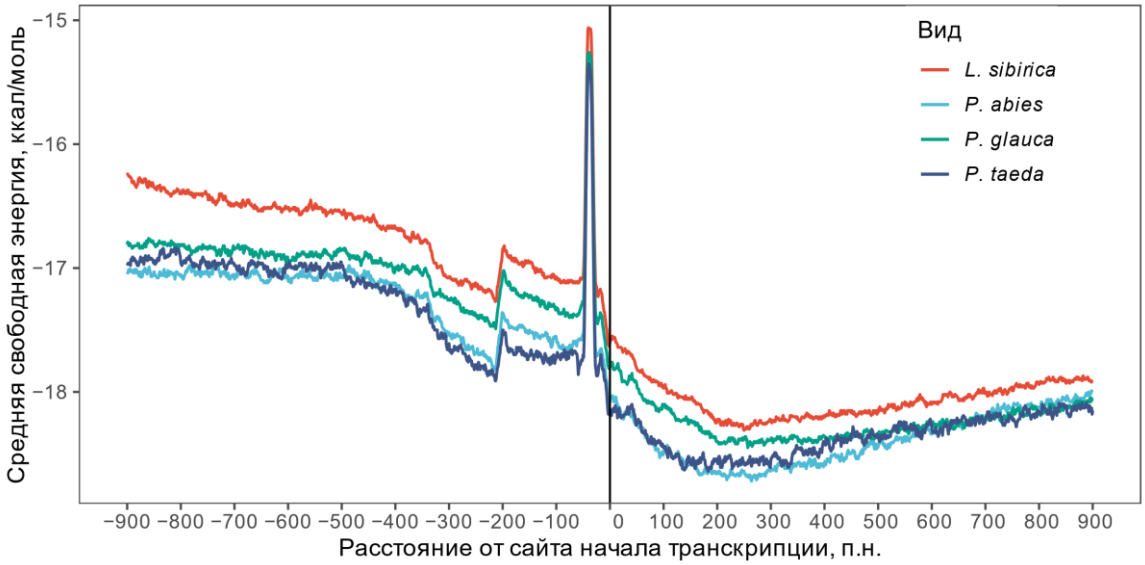


Рисунок 19. Распределение свободной энергии ДНК-дуплексов вокруг позиции TSS предсказанной TSSPlant.

Несколько курируемых баз данных содержат коллекции PWM, такие как JASPAR, PlantRegMap [338] или TRANSFAC [339]. Несмотря на постоянное совершенствование методов поиска PWM, они дают много ложноположительных предсказаний. Короткая длина и высокая вариабельность фактических мотивов связывания ТФ приводят к тому, что большинство совпадений с матрицей пропускают функциональные сайты связывания, либо предсказанные сайты имеют низкую вероятность в действительности быть функциональным [340]. Тем не менее, PWM-сканирование может служить полезным предварительным этапом для

создания списка TFBS-кандидатов, которые можно дополнительно отфильтровать с использованием других методов. Постоянство в расположении конкретных мотивов также может указывать на точно предсказанный промотор.

Чтобы определить позиционное распределение сайтов связывания транскрипциооных факторов (TFBS), мы просканировали идентифицированные промоторные области на наличие нескольких связанных с развитием и стрессом TFBS, используя TRANSFAC и MATCH. Позиционно-взвешенные матрицы (PWM), принадлежащие мотивам гомеодомена (Homeodomain), белкам теплового шока и транскрипционным факторам Myb, показывают два пика в их позиционном распределении (Рисунок 20b–d), в то время как мотивы связывания факторов AP2/EREBP имеют явное снижение вблизи области TATA-бокса (Рисунок 20) [204].

Надсемейство AP2/EREBP является одним из крупнейших ТФ растений и может быть разделено на три семейства: семейства AP2, RAV и ERF (фактор отклика на этилен). Наиболее распространенное семейство ERF делится на подсемейства ERF и DREB. ERF ТФ могут связываться с элементом GCC-box (AGCCGCC) и участвуют в путях передачи гормонов и регуляции генов, связанных с патогенезом. DREB ТФ связываются с элементом, чувствительным к обезвоживанию (DRE) с мотивом A/GCCGAC, и регулируют экспрессию генов, чувствительных к стрессу [341,342].

Гомеодоменовые ТФ представляют собой широко консервативные белки, составляющие примерно 15–30% всех ТФ у эукариот, которые управляют транскрипцией генов, ответственных за клеточную дифференцировку, морфогенез и поддержание плюрипотентности стволовых клеток. Они обладают ДНК-связывающим доменом, содержащим структуру спираль-поворот-спираль (helix-turn-helix, HTH), которая распознает короткий мотив 5'-TAAT-3' с очень умеренной специфичностью [343,344].

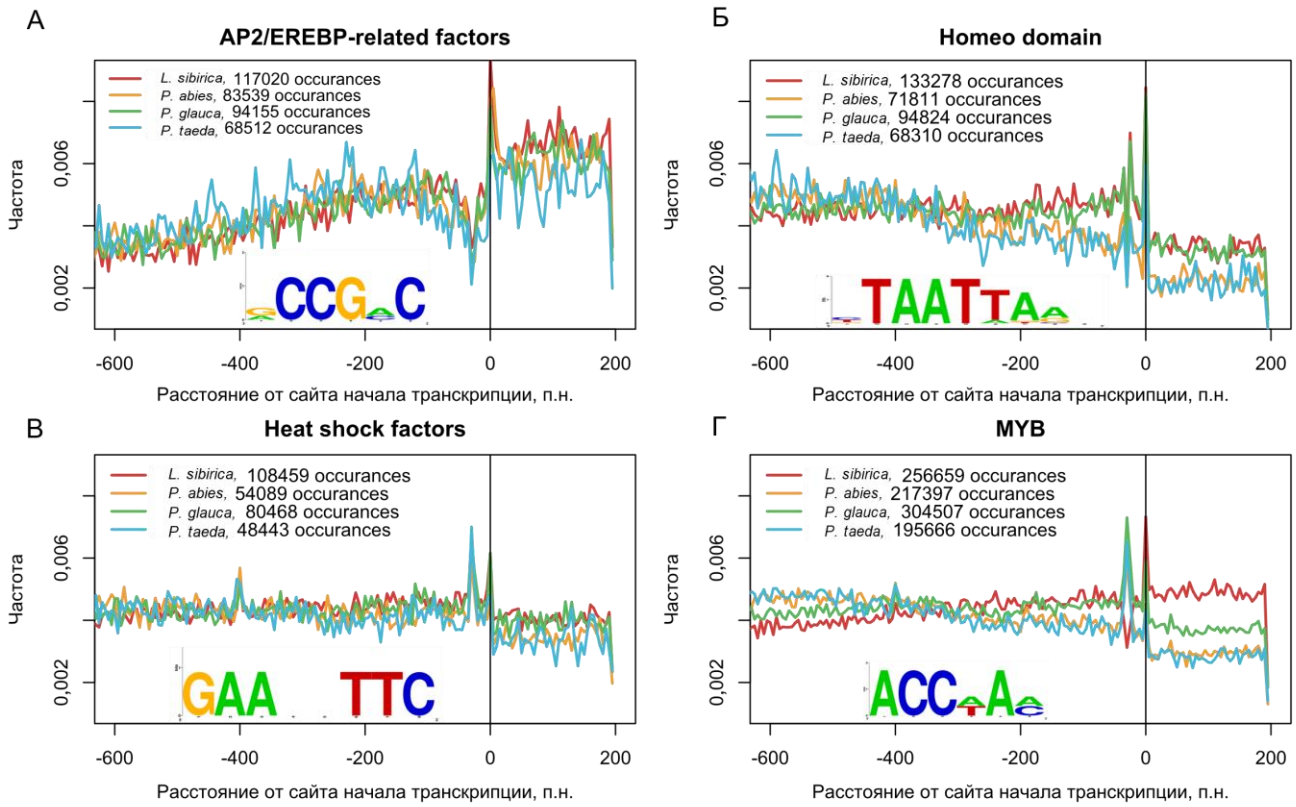


Рисунок 20. Позиционное распределение сайтов связывания факторов транскрипции (TFBS) для *Larix sibirica, Picea abies, Picea glauca* и *Pinus taeda* на основе PWM -сканирования с использованием TRANSFAC: **(a)** факторы, связанные с AP2/EREBP; (b) Homeodomain; **(c)** факторы транскрипции теплового шока; **(d)** Факторы транскрипции Myb.

Холод, соленость, засуха и другие стрессовые факторы, повреждающие белки, вызывают активацию и тримеризацию HSF, что позволяет связывать каждый мономер HSF с элементом теплового шока (HSE). HSE расположен в TSS генов HSP и включает по крайней мере два инвертированных повтора с консенсусным мотивом 5'-nGAAn-3' (5'-nGAAnnTTCn-3') выше TATA-бокса [345,346]. MYB-подобные белки контролируют метаболизм растений, развитие, судьбу клеток и реакцию на стресс. ТФ, содержащие характерный для растений домен MYB R2R3-типа, обычно связываются с AC-обогащенным мотивом ДНК (AC-элементами), таким как 5'-ACC(A/T)A(A/C)-3' [347,348]. PWM, принадлежащие Homeodomain, Heat shock и Myb ТФ, имеют два пика (Рисунок 20b–d), которые соответствуют AT-богатым участкам, на позициях 0 и –40 п.н., соответствующим TSS и TATA-box соответственно. Точно так же ТФ AP2 / EREBP с их GC- богатыми мотивами связывания имели выраженное падение вблизи области TATA-бокса (Рисунок 20a) [204].

3.3.2 Анализ нуклеотидного состава промоторов и кодирующих последовательностей

Большинство геномов покрытосеменных имеют отчетливый уменьшающийся от 5' до 3' градиент содержания GC в кодирующих областях. Этот эффект больше всего проявляется в третьем положении кодона. Существует возможная связь между рекомбинацией и градиентом 5'-3' GC, поскольку скорость рекомбинации выше вокруг TSS, что создает градиент рекомбинации 5'-3' [349,350]. Было высказано предположение, что градиент 5'-3' GC может указывать на инициацию рекомбинации на TSS [351].

Мы рассчитали GC_3 для всех кодирующих областей, которые имели поддержку RNA-seq. Подобно другим двудольным растениям, хвойные обладают унимодальным распределением GC_3 со средним значением 0,43 (sd = 0,087, Рисунок 21) [204]. Анализ кодирующих последовательностей у нескольких видов растений показал градиент GC_3 от 5'-конца к 3'-концу гена [137,352]. Все четыре проанализированных вида имели сходный градиент GC_3, который постепенно уменьшался начиная с 250 п.н. после старта кодирующей последовательности (Рисунок 21a).

Мы разделили гены на категории GC_3-бедных и GC_3-богатых, используя 10% и 90% квантили GC_3. Взаимосвязь между положением кодона в кодирующей последовательности и содержанием GC_3 была определена для обеих категорий GC_3, с помощью линейной регрессии на первых 1000 нуклеотидах кодирующей последовательности (Рисунок 21b). У лиственницы сибирской и риса наклон линии регрессии выражен сильнее, чем у сосны обыкновенной и арабидопсиса. Эти результаты согласуются с предыдущим сообщением о характере распределения GC в голосеменных растениях [140].

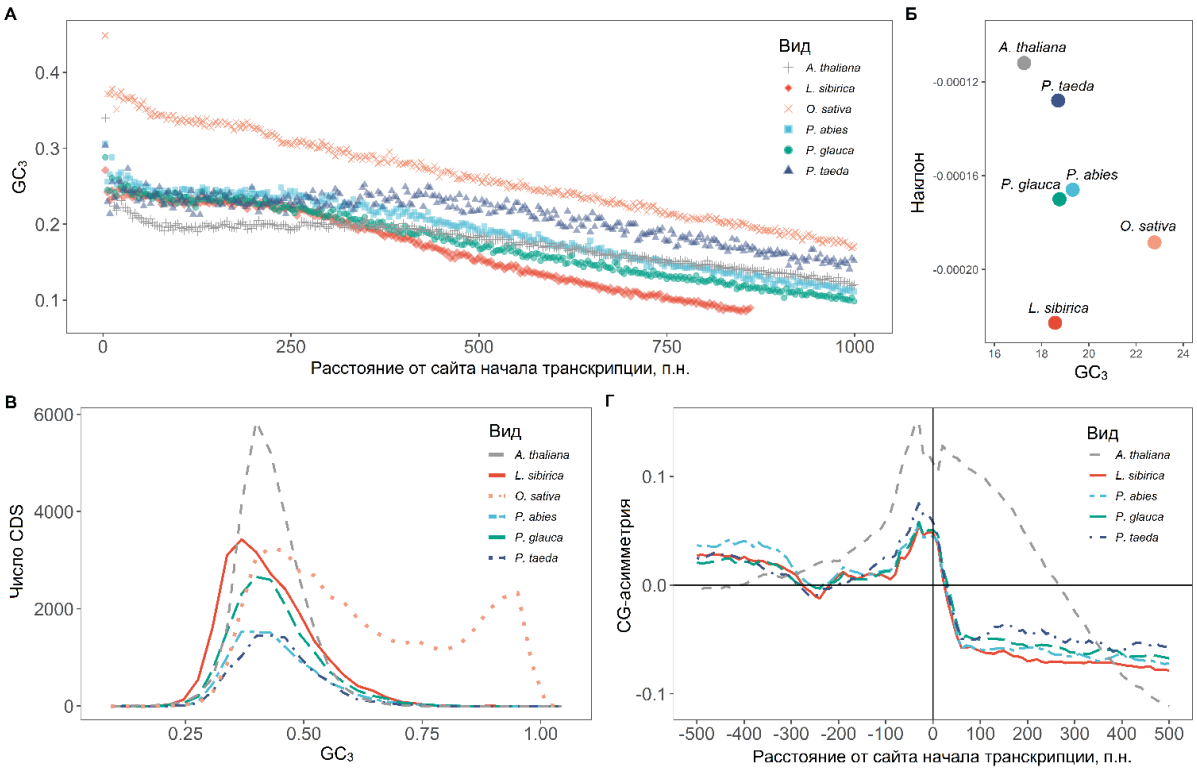


Рисунок 21. Некоторые статистические показатели распределения GC для четырех видов хвойных, *Larix sibirica, Picea abies, Picea glauca, Pinus taeda* и двух модельных видов растений, *Arabidopsis thaliana* и *Oryza sativa*: **А** — градиент GC_3 в кодирующих последовательностях, **Б** — наклон градиента GC_3, **В** — Распределение GC_3 по всем CDS, **Г** — CG-асимметрия вокруг TSS.

Обогащение ДНК гуаниновыми и цитозиновыми нуклеотидами связано с более высокой компактностью и плотностью генов и более высокими скоростями рекомбинации, чем менее обогащенные GC участки [351]. Было замечено, что у многих видов гены можно сгруппировать в два класса на основе содержания GC в третьем положении нуклеотидов кодирующих последовательностей [353–355]. Как сообщают Serres-Giardi с соавторами [140] и Татаринова с соавторами [137], у некоторых растений GC_3-бедные и GC_3-богатые гены значительно различаются по длине, при этом более длинные кодирующие последовательности, как правило, имеют более низкую частоту нуклеотидов G+C в третьем положении. Также считается, что преобладание нуклеотидов GC в более коротких генах является результатом их длины, так как содержание GC гена представляет собой среднее значение существующего градиента GC. Более короткие GC-богатые гены, как правило, либо моноэкзонны, либо имеют меньшее количество экзонов и интронов в целом, что напрямую влияет на среднее содержание GC, поскольку интроны

имеют более низкое содержание GC, чем экзоны. Согласно [351], унимодальное распределение содержания GC_3 указывает на меньший градиент GC внутри генов и более низкую скорость рекомбинации. Также ранее было замечено, что гены, богатые GC_3, могут демонстрировать более вариабельную экспрессию, чаще иметь TATA-зависимые промоторы и обычно участвуют в путях ответа на стресс. Наблюдаемый пик перекоса CG вокруг предсказанных TSS (Рисунок 21d) ранее был связан с эффективностью транскрипции и статусом метилирования генов, богатых GC_3 [138].

Подобно *A. thaliana* и *O. sativa* [134,135], GC-skew у четырех исследованных видов хвойных показал отчетливый пик вокруг TSS (Рисунок 21d). Высота пика у четырех видов хвойных ниже, чем у арабидопсиса. Это может быть связано с биологическими различиями или более низким качеством сборки генома [204].

Чтобы проверить, можно ли наблюдать разницу в длине генов между GC_3-бедными и GC_3-богатыми геномами в группе голосеменных, мы сравнили эти два класса генов (Рисунок 22). U-критерий Манна-Уитни показал, что длина CDS в генах с низким содержанием GC_3 была значительно больше, чем в генах с высоким содержанием GC_3 ($2{,}20 \times 10^{-16} < p < 6{,}09 \times 10^{-12}$) [204].

Между двумя классами генов была значительная разница в количестве экзонов; эта тенденция сохраняется для всех исследованных голосеменных и покрытосеменных растений. Согласно текущим аннотациям генома, GC_3-богатые гены, как правило, имеют от двух до четырех экзонов у лиственницы сибирской и сосны обыкновенной и два или меньше экзона у европейской и белой ели (Рисунок 23) [204]. Гены с более чем пятью экзонами, как правило, являются GC_3-бедными у всех видов.

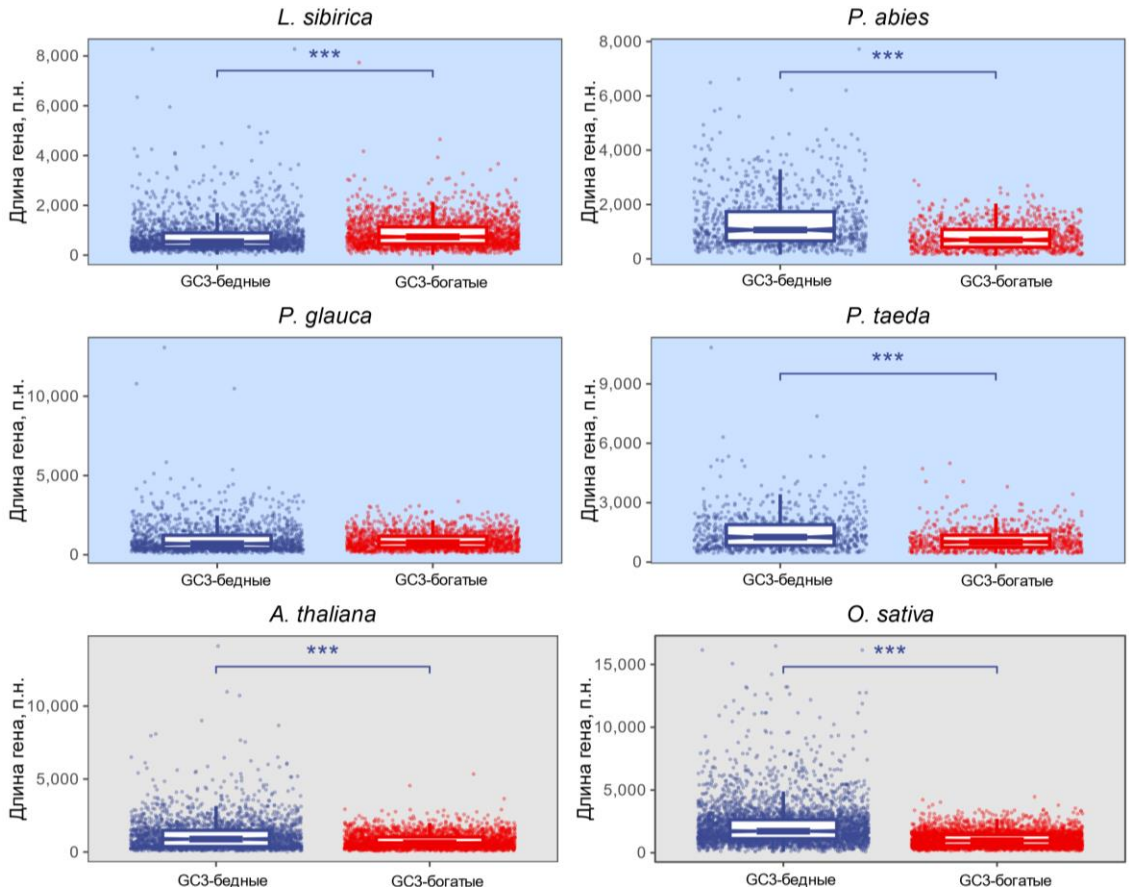


Рисунок 22. Разница в длине кодирующей последовательности между GC3-бедными и GC3-богатыми генами. 10% и 90% квантили использовались для разделения генов на классы с низким содержанием GC3 и с высоким содержанием GC3 (синий и красный соответственно).

CG-skew в геномах четырех хвойных были ниже, чем у *A. thaliana*, но чтобы сделать вывод, было ли это связано с качеством сборки генома хвойных или с различиями между голосеменными и покрытосеменными (или однодольными и двудольными), необходимо проанализировать гораздо больше геномов, что выходит за рамки данного исследования [204]. Тем не менее, это можно сделать с помощью представленных здесь инструментов, и это может стать перспективным направлением для будущих исследований.

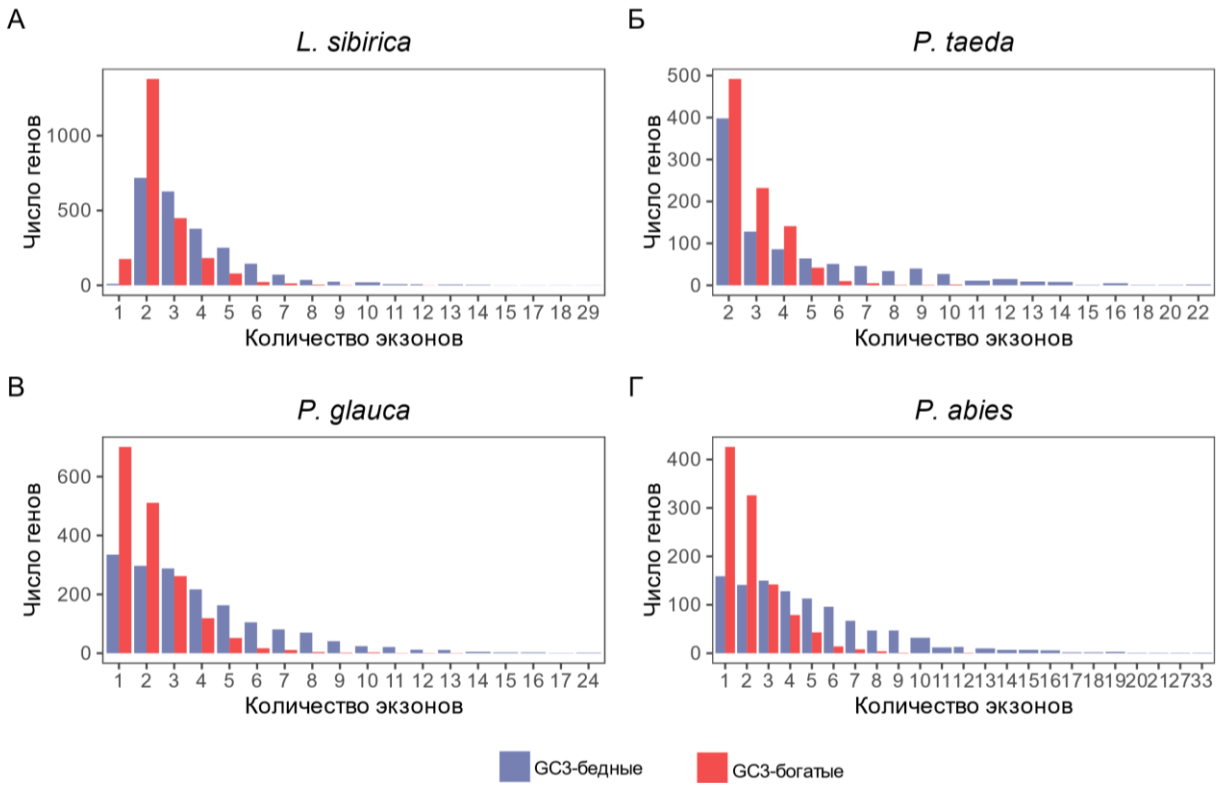


Рисунок 23. Распределение числа экзонов на ген в GC3-бедных и GC3-богатых генах у *L. sibirica, P. abies, P. glauca* и *P. taeda*. Количество генов в категориях GC3-бедных и GC3-богатых было одинаковым в каждом организме.

3.4 Разработка микросателлитных маркеров для оценки генетического разнообразия лиственниц сибирской, Гмелина и Каяндера

3.4.1 Отбор повторов и дизайн праймеров

После работы алгоритма по поиску тандемных повторов GMATo было найдено 1015 регионов с повторяющимися трех-, четырех-, пяти- и шестинуклеотидными мотивами (Таблица 9). После отбора тандемных повторов по минимальному числу повторения мотивов и положению в контиге при помощи WebSat был выполнен дизайн праймеров для выбранных 222 микросателлитных локусов (Таблица 9) [30].

Таблица 9 — Микросателлитные маркеры для ядреного генома *L. sibirica*, отобранные для дизайна праймеров

Нуклеотидный мотив	**Минимальное число повтора мотива**	**Число соответствующих SSR-повторов**	**Число микросателлитов с праймерами**	**Количество пар праймеров после проверки на специфичность**
Трехнуклеотидные	15	423	78	27
Четырехнуклеотидные	10	249	57	12
Пятинуклеотидные	7	132	36	12
Шестинуклеотидные	7	211	51	9
		∑ = 1015	∑ = 222	∑ = 60

Выравнивание контигов, содержащих отобранные повторы, на последовательности митохондрий растений выявило наличие среди отобранных контигов одного, вероятнее всего принадлежащего митохондриальному геному. Данный контиг с соответствующим ему микросателлитным повтором из дальнейшего анализа был исключен в данной работе, но может представлять собой ценный маркер для изучения материнский линий. Проверка на наличие хлоропластных последовательностей не выявила их присутствия среди отобранных контигов [30].

После проверки полученных последовательностей праймеров на специфичность при помощи выравнивания на исходную сборку была выявлена 161 пара праймеров, имеющих более одного точного совпадения в использованной сборке ядерного генома. Эти праймеры были признаны недостаточно специфичными для амплификации конкретных SSR-локусов, и удалены из дальнейшего анализа. Оставшиеся 60 пар праймеров были отобраны для дальнейшего тестирования в лаборатории (Приложение Д.1) [30].

3.4.2 Отбор полиморфных маркеров

Первичный анализ отобранных праймеров показал, что из 60 пар праймеров 14 обнаружили отсутствие амплифицированного продукта, а для остальных 13 амплифицированный продукт был не того размера. Остальные 33 локуса проявили более-менее устойчивую амплификацию в подобранных условиях. Эти праймеры

далее тестировались на увеличенной выборке деревьев. По результатам данной проверки было отобрано 23 локуса, демонстрировавших наиболее стабильные интерпретируемые спектры [30].

Для уточнения полиморфности маркеров была проведена проверка всех 23 маркеров на выборках от трех видов из трех географически отдаленных популяций — *L. sibirica*, *L. gmelinii* и *L. cajanderi*. Два локуса, *Ls_8778872* и *Ls_291006*, были признаны мономорфными для трех видов лиственницы и исключены из дальнейшего анализа. Еще четыре локуса, *Ls_2672894, Ls_4040657, Ls_980491* и *Ls_3542003*, при тестировании на 8 образцах показали себя мономорфными для *L. sibirica*, но полиморфными для *L. gmelinii* и *L. cajanderi*. Два локуса, *Ls_3765334* и *Ls_254200*, проявили себя как мономорфные для *L. gmelinii* и *L. cajanderi,* но полиморфные для *L. sibirica* (Таблица 10**)** [30].

Таблица 10 — Результаты тестирования 23 ядерных микросателлитных локусов на 8 образцах.

Локус	**Мотив**	**Размер продукта, п.н.**	***L. sibirica***	***L. gmelinii***	***L. cajanderi***
Ls_417667	AAT	238	П	П	П
Ls_840190	TAC	232	П	П	П
Ls_951631	ATC	150	П	П	П
Ls_954234	ATT	202	П	П	П
Ls_611965	CAG	236	П	П	П
Ls_752897	AAG	252	П	П	П
Ls_1247092(2)	CTT	214	П	П	П
Ls_3765334	GAG	270	П	М	М
Ls_254200	AAT	252	П	М	М
Ls_1664757	TCT	144	П	П	П
Ls_3542003	TCA	142	М	П	П
Ls_291006	AAAG	168	М	М	М
Ls_1008427	ATAG	181	П	П	П
Ls_1524449	ATAG	179	П	П	П
Ls_2672894	TTTG	163	М	П	П
Ls_2552367	CTAT	280	П	П	П
Ls_980491	CTAT	228	М	П	П
Ls_1898261	ACAT	191	П	П	П
Ls_3952800	TATG	252	П	П	П
Ls_897755	ACCAT	265	П	П	П
Ls_305132	GTCGGA	219	П	П	П
Ls_8778872	TGTTGA	221	М	М	М
Ls_4040657	TCACTT	245	М	П	П

Примечание. П — полиморфный, М — мономорфный

Тестирование отобранного 21 локуса на 24 образцах из четырех географически отдаленных популяций лиственницы показало, что при расширении выборки локусы *Ls_1664757* и *Ls_3542003* по-прежнему остаются мономорфными (Рисунок 24).

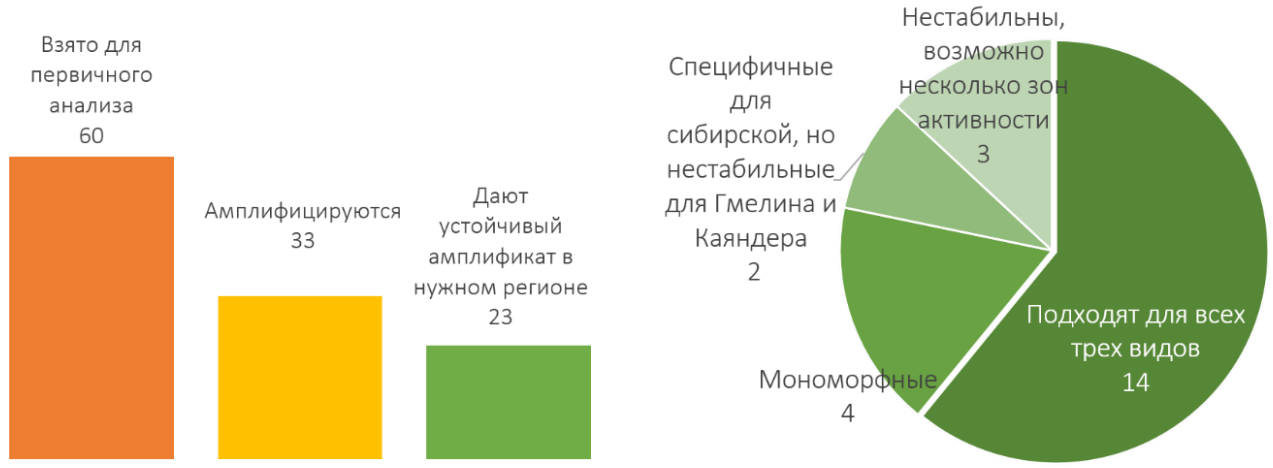


Рисунок 24. Число микросателлитных локусов на каждом из этапов проверки разработанных праймеров.

Остальные 14 протестированных локусов были полиморфными и пригодными для проведения пробного популяционно-генетического анализа (Приложение Д.2). Из них наиболее простые спектры с одной зоной активности и числом аллелей от 3 до 5 показали семь локусов (*Ls_980491*, *Ls_2672894*, *Ls_4040657*, *Ls_1008427*, *Ls_417667*, *Ls_2552367* и *Ls_3765334*). Примеры электрофореграмм данных локусов показаны на Рисунке 25.

Локус *Ls_1524449* является высокополиморфным для *L. sibirica*, с 9 аллельными вариантами, тогда как для *L. gmelinii* и *L. cajanderi* данный маркер показал нестабильные спектры с большим числом нуль-аллелей, с отсутствием амплификации, вероятно, вследствие мутаций в зоне отжига праймеров. Локус *Ls_951631* также показал себя полиморфным для *L. sibirica*, но в популяции лиственницы Каяндера дает слишком большой разброс в различии аллельных вариантов и, вероятно, имеет две перекрывающихся зоны активности, что сильно затрудняет генотипирование. Локус *Ls_254200*, при первичной проверке выглядевший мономорфным на лиственницах Каяндера и Гмелина, в дальнейшем проявил слабую полиморфность для этих видов, а для *L. sibirica* был полиморфным,

но с большим числом нуль-аллелей и нестабильной амплификацией. Для локусов *Ls_1898261* и *Ls_897755* обнаружено слишком большое количество неспецифической амплификации [30].

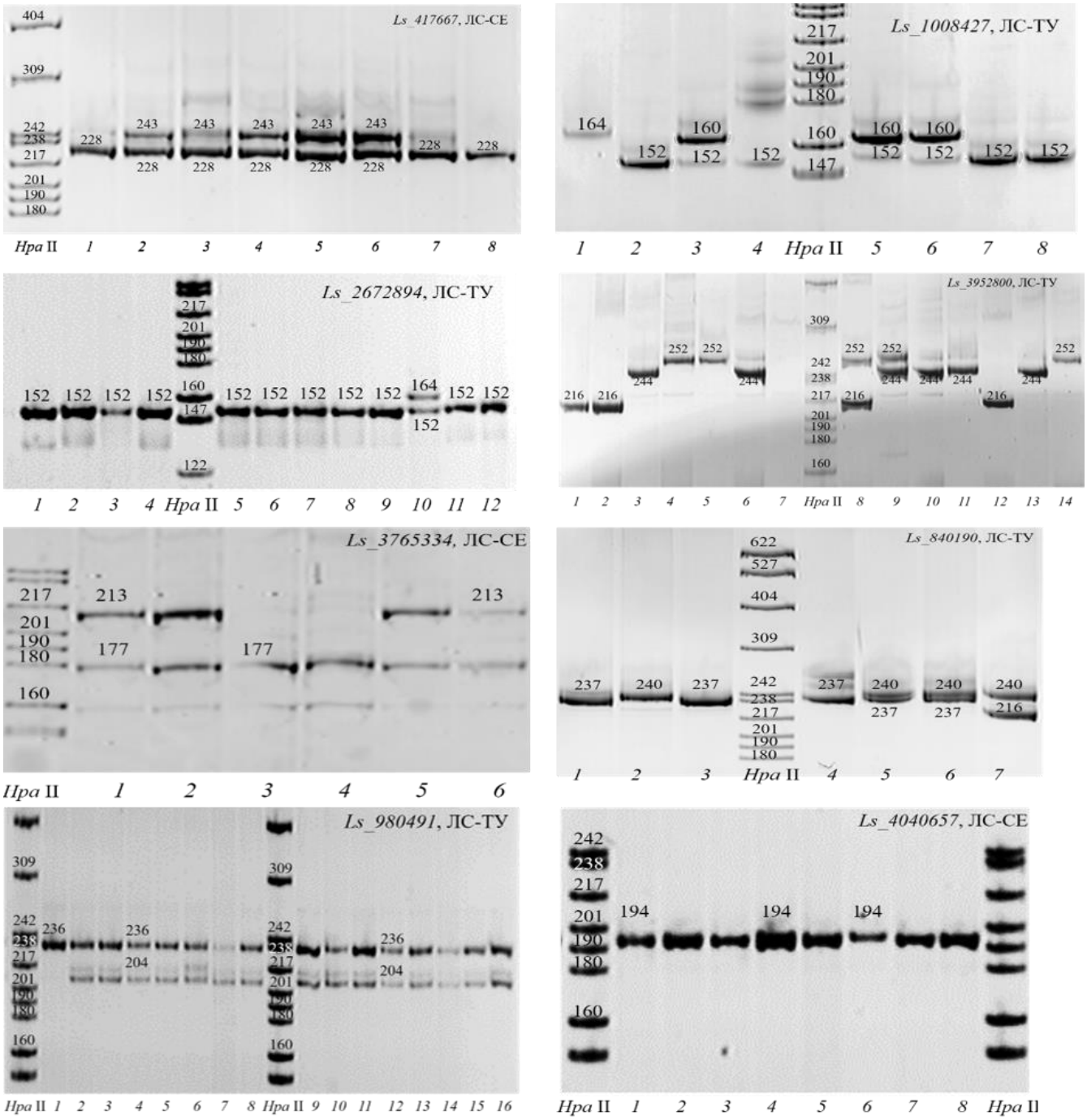


Рисунок 25. Примеры электрофореграмм некотрых ядерных микросателлитных локусов, цифрами обозначены аллели по длине фрагмента ДНК (лист 1 из 2, фото автора).

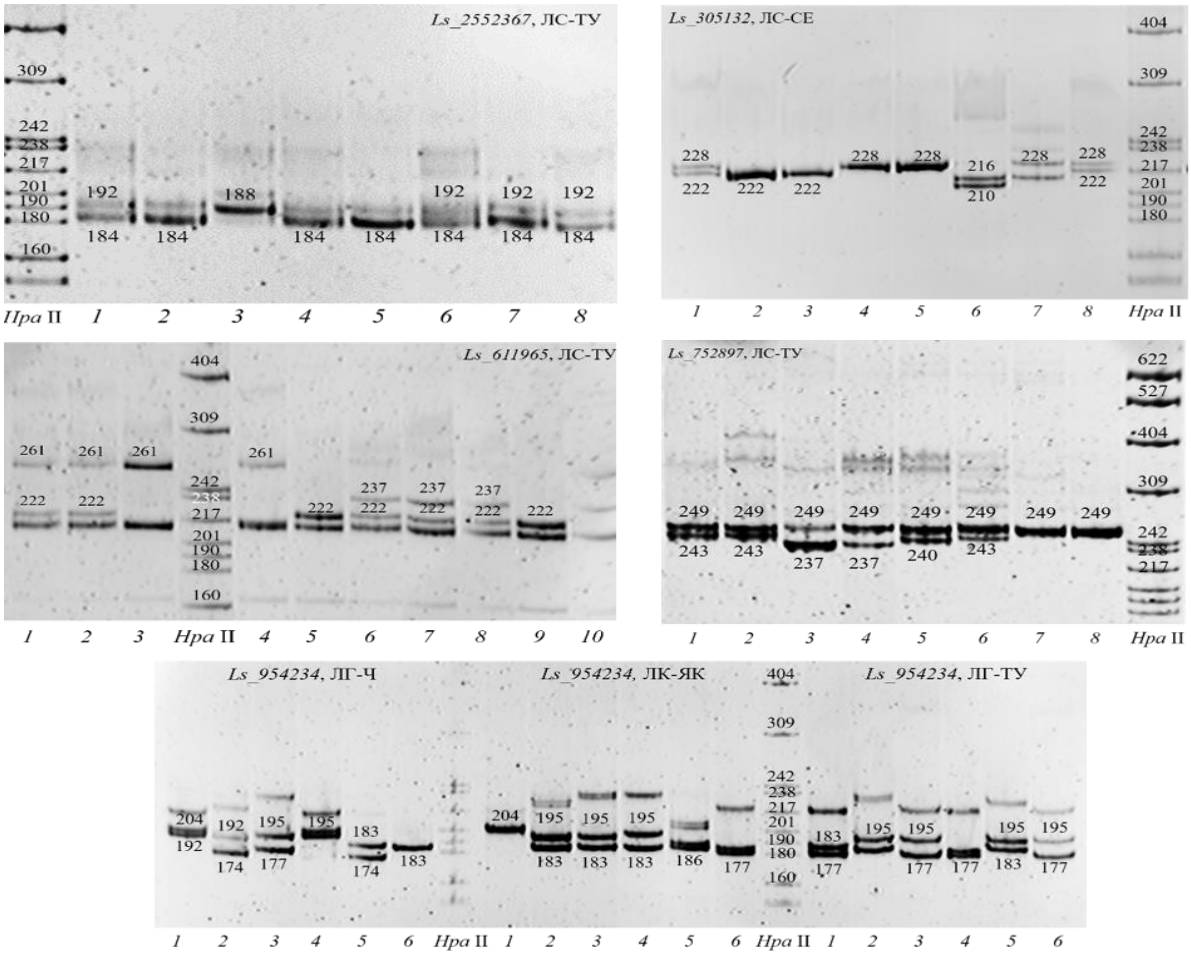


Рисунок 25 (лист 2 из 2, фото автора).

3.4.3 Оценка показателей генетического разнообразия видов *L. sibirica, L. gmelinii* и *L. cajanderi*

В процессе исследования 14 ядерных микросателлитных локусов в четырех популяциях лиственниц сибирской, Гмелина и Каяндера из различных районов их естественного произрастания выявлено 86 аллельных вариантов, 28 (32,5%) из которых оказались общими. Так же для 9 локусов были выявлены нуль-аллели. У изученных нами популяций идентифицированные микросателлитные локусы существенно отличались по составу и частотам встречаемости выявленных аллелей. Так, например, у лиственницы сибирской выявлено 57 аллелей, из которых 11 оказались специфичными. У лиственницы Гмелина выявлено 62 аллеля, из которых только 7 специфичные, а у лиственницы Каяндера — 61 аллель и специфичных лишь 5. Локус *Ls_980491* в популяции лиственницы из Забайкалья проявил себя как мономорфный. Так же мономорфным себя показал и локус *Ls_4040657* для обеих выборок лиственниц сибирских (Приложение Д.3) [30].

Самый высокий уровень аллельного разнообразия в исследованных популяциях лиственницы имеют локусы *Ls_752897, Ls_954234 Ls_3952800,* в которых выявлено 13, 11 и 10 аллелей соответственно (Приложение Д.3).

На основании аллельных частот 14 локусов были рассчитаны основные показатели генетической изменчивости исследованных популяций лиственниц сибирской, Гмелина и Каяндера (Таблица 11).

Таблица 11 — Показатели генетической изменчивости, рассчитанные для трех видов лиственниц по результатам SSR-анализа

Популяция	***N***	N_A	N_E	H_O	H_E	***F***
Larix sibirica						
ЛС-СЕ	24	4,07 ± 0,47	2,39 ± 0,25	0,56 ± 0,07	0,50 ± 0,06	–0,12 ± 0,07
ЛС-ТУ	24	3,29 ± 0,32	2,22 ± 0,22	0,56 ± 0,08	0,48 ± 0,06	–0,16 ± 0,09
В среднем для вида		3,68 ± 0,29	2,31 ± 0,16	0,56 ± 0,05	0,49 ± 0,04	–0,14 ± 0,06
Larix gmelinii						
ЛГ-Ч	24	4,9 ± 0,7	2,7 ± 0,52	0,46 ± 0,08	0,47 ± 0,07	0,01 ± 0,03
Larix cajanderi						
ЛК-Я	24	4,50 ± 0,61	2,35 ± 0,43	0,41 ± 0,07	0,43 ± 0,07	0,06 ± 0,05
В среднем для всех популяций		4,19 ± 0,28	2,41 ± 0,18	0,49 ± 0,04	0,47 ± 0,03	–0,05 ± 0,03

Примечание. N — число деревьев в выборке, N_A — среднее число аллелей на локус, N_E — эффективное число аллелей на локус, H_O — наблюдаемая гетерозиготность, H_E — ожидаемая гетерозиготность, *F*— индекс фиксации, ± — стандартная ошибка.

Анализ показателей генетического разнообразия показал, что наиболее высокое аллельное разнообразие было выявлено в популяциях лиственниц Гмелина (N_A = 4,93 и N_E = 2,69) и Каяндера (N_A = 4,50 и N_E = 2,35), что объясняется наличием большего количества редких аллелей, чем у лиственницы сибирской. Однако по средним уровням наблюдаемой и ожидаемой гетерозиготности гораздо большую изменчивость показывают выборки из популяций лиственницы сибирской (H_O = 0,558 и H_E = 0,491) [30]. Такая картина вполне согласуется с литературными данными [28], и может быть объяснена условиями, в которых произрастают данные виды.

В изученных популяциях лиственниц Гмелина и Каяндера выявлен дефицит гетерозиготных генотипов, тогда как в двух популяциях лиственницы сибирской, напротив, обнаружен небольшой избыток гетерозигот (*F* = –0,142). Высокие индексы фиксации были выявлены в якутской и читинской выборках лиственниц

Каяндера (F = 0,061) и Гмелина (F = 0,009). Это может быть связано с инбридингом или подразделённостью популяций. Особенности этих выборок, приведшие к такому высокому значению индекса фиксации (F), заключается в том, что данные регионы регулярно подвергаются сильным пожарам. Послепожарное восстановление древостоя на данных участках происходило главным образом за счет небольшого числа неповрежденных или слабо повреждённых деревьев, что может вести к инбридингу. Но инбридинг оказывает одинаковое влияние на все локусы, и положительные значения индекса должны наблюдаться для всех или большинства локусов. Однако они наблюдаются только для некоторых локусов, что может объясняться не инбридингом, а подразделённостью популяции или наличием нуль-аллелей в данных локусах [30]. В целом, в выборках лиственницы из Якутии и читинского района избыток гомозигот подтверждают ранее полученные аналогичные данные в северных популяциях рода *Larix* [356]. Высокий уровень инбридинга характерен для видов рода *Larix* по причине высокой частоты самоопыления, поскольку пыльца лиственницы не имеет воздушных мешков, в следствии чего не летит от кроны на большое расстояние [28,356–358].

Оценка популяционной структуры на основе F-статистик Райта (Таблица 12) показала, что индекс фиксации особи относительно популяции в среднем составляет около 8% (F_{IS} = –0,077), и относительно вида 7% (F_{IT} = 0,074) [211]. Приблизительно 13% от всей наблюдаемой изменчивости (F_{ST} = 0,134) приходится на межпопуляционную. Внутри популяций сосредоточено 86,6% всего генетического разнообразия. Наибольший вклад в дифференциацию изученных популяций вносят локусы *Ls_980491, Ls_611965, Ls_840190* и *Ls_3765334* (Таблица 12). Данные локусы могут служить в качестве диагностических для генетической маркировки выборок из разных географических районов [30]. Приведенные оценки межпопуляционной изменчивости согласуются с ранее опубликованными данными, как по анализу SSR-маркеров [28], так и по анализу аллозимных маркеров для лиственниц сибирской и Гмелина [19,356].

Таблица 12 — Значение показателей *F*-статистик Райта

Локус	Число аллелей	F_{IS}	F_{IT}	F_{ST}	Значения $\chi 2$
Ls_1008427	5	0,17	0,26	0,10	76,11 (15) ***
Ls_1247092(2)	7	-0,09	0,03	0,10	106,14 (28) ***
Ls_2552367	4	-0,01	0,03	0,04	15,78 (6) *
Ls_2672894	3	-0,08	-0,05	0,03	0,43 (3) ns
Ls_305132	6	-0,02	0,12	0,14	28,85 (21) ns
Ls_3765334	5	-0,11	0,08	0,17	27,71 (15) *
Ls_3952800	9	-0,13	0,04	0,15	253,67 (45) ***
Ls_4040657	3	0,14	0,22	0,09	65,23 (6) ***
Ls_417667	5	-0,49	-0,26	0,16	11,94 (10) ns
Ls_611965	7	0,08	0,32	0,26	38,87 (21) *
Ls_752897	12	0,07	0,18	0,12	210,36 (78) ***
Ls_840190	7	0,19	0,34	0,18	59,71 (28) ***
Ls_954234	10	-0,07	-0,01	0,06	146,08 (55) ***
Ls_980491	3	-0,74	-0,27	0,27	8,56 (3) *
Среднее		**-0,077 ± 0,068**	**0,074 ± 0,050**	**0,134 ± 0,020**	

Примечание: F_{IS} — индекс фиксации особи относительно популяции, F_{IT} — индекс фиксации особи относительно вида, F_{ST} — коэффициент межпопуляционной дифференциации. $\chi 2$ — тест на гетерогенность аллельных частот; в скобках указано число степеней свободы. Уровни значимости: * — $P < 0,05$; ** — $P < 0,01$; *** — $P < 0,001$, ns — не значимо.

Уровень генетической дифференциации между исследованными популяциями был определён с использованием стандартного генетического расстояния (D_N) Нея [212], на основании частот аллелей 14 SSR-локусов (Таблица 13).

Таблица 13 — Генетические расстояния D_N между изученными популяциями лиственниц.

	ЛГ-Ч	**ЛК-Я**	**ЛС-СЕ**	**ЛС-ТУ**
ЛГ-Ч	–			
ЛК-Я	0,104	–		
ЛС-СЕ	0,246	0,282	–	
ЛС-ТУ	0,265	0,323	0,092	–

Из приведенных в Таблице 13 данных видно, что значения D_N между популяциями лиственницы варьируют в достаточно широких пределах: от 0,09 до 0,32. Анализ индивидуальных генотипов (генотипических дистанций) [213] изученных видов лиственницы также показал подразделенность популяций и соответствие их географическому расположению (Рисунок 26) [30].

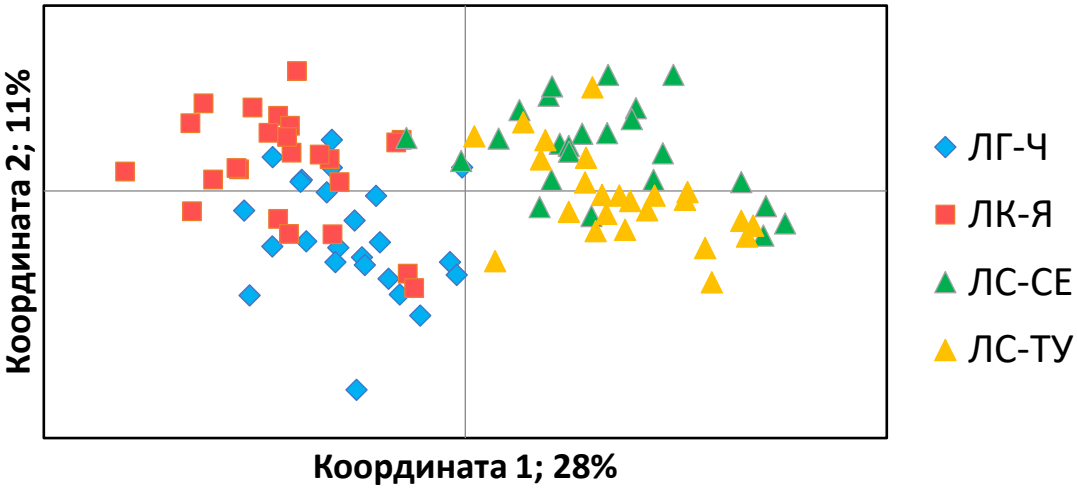


Рисунок 26. Проекция изученных выборок лиственницы на плоскости двух координат по данным анализа главных координат (PCoA) матрицы генотипических расстояний.

Согласно данным расчётов, наиболее генетически удалёнными друг от друга являются выборки лиственницы Каяндера из Якутии и лиственницы сибирской из Хакасии (D_N = 0,323). Наименьшее значение генетического расстояния наблюдается между выборками одного вида — лиственницы сибирской (D_N = 0,092), что вполне естественно. Стоит так же обратить внимание, что значение расстояния Нея между выборками из лиственниц Гмелина и Каяндера лишь немногим больше (D_N = 0,104) [30]. Эти данные могут служить подтверждением гипотезы о том, что данные виды *L. gmelinii* и *L. cajanderi* на самом деле являются не разными видами, а, скорее, двумя расами одного вида [19,359]. Установленный уровень дифференциации включенных в исследование выборок из популяций лиственницы наглядно показывает расположение популяций на плоскости двух координат (Рисунок 26).

Таким образом, использование 14 ядерных микросателлитных локусов, разработанных в данной работе, позволило получить оценки генетического разнообразия и дифференциации для выборок из популяций широко распространенных видов лиственниц сибирской, Гмелина и Каяндера из разных частей их ареалов. Полученные результаты согласуются с аналогичными данными, полученными не только с использованием SSR-маркеров, разработанных для других видов лиственниц [28,360], но и с данными, полученными по другим генетическим маркерам для рода *Larix* [19,156,356,358,360].

ЗАКЛЮЧЕНИЕ

Работа с полными геномами хвойных представляет собой нетривиальную задачу, так как из-за огромных размеров требует серьезных вычислительных ресурсов (время вычислений, объем памяти), необходимых для обработки геномных данных. Структурная аннотация полногеномной сборки лиственницы сибирской на 448 ядрах вычислительного кластера заняла 22 дня, а отдельный запуск RepeatMasker для идентификации повторов на основе гомологии на полной сборке генома с использованием 40 ядер занял 20 дней.

В рамках данной работы впервые получена подробная структурная и функциональная аннотация генов для лиственницы сибирской, а также были получены сборки транскриптомов нескольких тканей. Эти данные представляют собой первый публично доступный ресурс для рода *Larix*. Сравнение полученных в ходе аннотации белок-кодирующих генов с набором генов арабидопсиса показывает, что вероятно большая часть генов (72%) была идентифицирована и охарактеризована. Таким образом, данную аннотацию можно использовать в качестве ресурса и основного референса для дальнейших геномных исследований рода *Larix*. Несмотря на фрагментарность и неполноту, сборки и аннотации геномов хвойных по-прежнему представляют собой ценный ресурс для дальнейших геномных и генетических исследований. Текущее состояние геномных аннотаций хвойных позволяет сравнивать различия между видами голосеменных и покрытосеменных растений на геномном уровне, что было продемонстрировано в данной работе на примере различий в представленности генов в различных функциональных категориях, таких как организация клеточной стенки и метаболизм, программируемая клеточная смерть и аутофагия, биосинтез гормонов стресса.

Характерной особенностью геномов хвойных является большое количество повторов, в том числе транспозонов и ретротранспозонов. Типы выявленных повторов и их распределение в геноме лиственницы сибирской соответствуют таковым у других хвойных. Доля генома, покрытого повторами в порции длинных прочтений Oxford Nanopore, по оценке RepeatMasker, составила 66%, что близко к

таковым оценкам для лиственниц японской, однако свидетельствует о том, что часть повторов в геноме лиственницы сибирской была слишком фрагментирована, чтобы быть включенной в окончательную сборку. В данной работе была получена комплексная видоспецифичная библиотека повторов, которая может использоваться для поиска мобильных элементов в геномах других голосеменных растений.

На основании проведенной идентификации LTR, а также имеющихся в литературе данных, вероятный период массированного встраивания ретротранспозонов в геном лиственницы произошел порядка 4-5 млн лет назад. Типичные оценки времени встраивания LTR в геномы растений варьируются от 1 до 2,5 млн лет назад для покрытосеменных растений и 10–15 млн лет назад для голосеменных. На данную оценку у лиственницы могут влиять как эффективный механизм элиминации повторов в сочетании с вставкой древних повторов, так и фрагментарный характер черновой сборки, приведший к малому количеству найденных LTR.

Для трёх других видов семейства *Pinaceae* были предсказаны сайты начала транскрипции, с помощью вычислительных подходов, основанных на методе максимизации ожидания и классификации нейронной сетью. Был опробован метод валидации предсказаний *de novo* на основе распределения длин 5′-нетранслируемой области, профиля распределения свободной энергии ДНК дуплексов и позиционного распределения сайтов связывания транскрипционных факторов. Эта работа является первым полногеномным предсказанием TSS в геномах размером более 10 млрд.п.н. Предсказанные TSS и соответствующие им промоторные области обеспечивают основу для будущей экспериментальной проверки и представляют собой ценный ресурс для лучшего понимания регуляции генов и исследования эволюционных отношений между голосеменными и покрытосеменными. Идентификация TSS может найти свое применение в генетической селекции и редактировании генома, предоставляя возможности для более точного картирования в функциональных областях генома и локусах количественных признаков, связанных с адаптивными чертами, такими как

скорость роста, устойчивость к холоду и засухе. резистентность и устойчивость к инвазии патогенов.

На основе полногеномных данных были разработаны и проверены 14 перспективных микросателлитных локусов для лиственницы сибирской, демонстрирующих также средне- и высоко-полиморфные спектры для лиственниц Гмелина и Каяндера. Результаты первичного популяционно-генетического анализа, проведенного с использованием разработанных SSR-маркеров, позволили получить оценки уровня генетического разнообразия и дифференциации четырех выборок из популяций лиственниц сибирской, Гмелина и Каяндера. Разработанные в данной работе маркеры могут успешно применяться для изучения не только лиственницы сибирской, но также лиственниц Гмелина и Каяндера. Дальнейший анализ уровней изменчивости природных и искусственных популяций лиственницы с помощью предложенных маркеров позволит получить количественные оценки их генетической структуры, таких как внутрипопуляционное аллельное и генное разнообразие, генетическая подразделенность и дифференциация на разных иерархических уровнях, степень инбридинга и др.

ВЫВОДЫ

1. Впервые получена подробная структурная и функциональная аннотация ядерного, митохондриального и хлоропластного геномов, а также сборки транскриптомов нескольких тканей для лиственницы сибирской. Эти данные представляют собой первый публично доступный геномный ресурс для рода *Larix*.

2. На примере поиска различий в представленности генов в функциональных категориях организации клеточной стенки и метаболизма, программируемой клеточной смерти и аутофагии и биосинтеза гормонов стресса, продемонстрирована возможность использования данной аннотации, а также аннотаций других видов семейства Pinaceae в качестве основного референса для сравнительного геномного анализа.

3. Оценка доли повторов в геноме лиственницы на основе длинных прочтений Oxford Nanopore составила 66%, что позволяет предположить, что репитомная часть генома была слишком фрагментирована, чтобы быть включенной в окончательную сборку. Получена комплексная видоспецифичная библиотека повторов, которая может использоваться для поиска мобильных элементов в геномах других голосеменных растений.

4. На основании идентифицированных интактных длинных концевых повторов, а также имеющихся в литературе данных, вероятный период массированного встраивания ретротранспозонов в геном лиственницы может быть оценен порядка 4-5 млн лет назад.

5. Для *L. sibirica*, а также для трех других видов семейства Pinaceae — *P. abies, P. glauca* и *P. taeda* — были предсказаны сайты начала транскрипции. Был опробован метод валидации предсказаний *de novo* на основе распределения длин 5’-UTR, профиля распределения свободной энергии ДНК и позиционного распределения сайтов связывания транскрипционных факторов. Эта работа является первым полногеномным предсказанием TSS в геномах размером более 10 млрд.п.н.

6. Были разработаны 14 микросателлитных маркеров для лиственницы сибирской, демонстрирующих также средне- и высоко-полиморфные спектры для лиственниц Гмелина и Каяндера и удобные для анализа простым гель-электрофорезом в лабраториях не оборудованных дорогостоящими приборами для капиллярного гель-электрофореза. Результаты первичного популяционно-генетического анализа, проведенного с использованием разработанных маркеров, позволили получить оценки уровня генетического разнообразия и дифференциации четырех выборок из популяций лиственниц сибирской, Гмелина и Каяндера.

Список сокращений и условных обозначений

АБК — абсцизовая кислота
ДНК — дезоксирибонуклеиновая кислота
ЖАК — жасмоновая кислота
млн.л.н. — миллион лет назад
млн.т.п. — миллионов пар нуклеотидов
млрд.п.н. — миллиардов пар нуклеотидов
п.н. — пар нуклеотидов
РНК — рибонуклеиновая кислота
т.п.н. — тысяч пар нуклеотидов
тыс.л.н. — тысяч лет назад
ТФ — транскрипционный фактор
BLAST — basic local alignment search tool
BUSCO — benchmarking universal single-copy orthologs
EST — expressed sequence tag
FDR — false discovery rate
GO — gene ontology
LRR — leucine-rich repeat
LTR — long tandem repeat
MITE — miniature inverted-repeat transposable elements
NCBI — national center for biotechnology information
NGS — next generation sequencing
PWM — positional weight matrix
UTR — untranslated region
SNP — single-nucleotide polymorphism
SSR — simple sequence repeat
TAIR — the Arabidopsis Information Resource
TE — transposable element
TFBS — transcription factor binding site
TIR — terminal inverted repeat
TSA — transcriptome shotgun assembly
TSD — target site duplication
TSS — transcription start site
WGD — whole genome duplication

Список литературы

1. Kress, W.J. Green plant genomes: What we know in an era of rapidly expanding opportunities / W.J. Kress et al. // Proc. Natl. Acad. Sci. U.S.A. – 2022. – Vol. 119, № 4. – P. e2115640118.
2. NCBI Genome // Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. – 2023. URL: https://www.ncbi.nlm.nih.gov/genome/ (accessed: 06.06.2023).
3. Lughadha, E.N. Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates: 1 / E.N. Lughadha et al. // Phytotaxa. – 2016. – Vol. 272, № 1. – P. 82-88.
4. Qiao, X. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants / X. Qiao et al. // Genome Biology. – 2019. – Vol. 20, № 1. – P. 38.
5. Pellicer, J. Genome Size Diversity and Its Impact on the Evolution of Land Plants / J.Pellicer et al // Genes (Basel). – 2018. – Vol. 9, № 2. – P. 88.
6. Claros, M.G. Why Assembling Plant Genome Sequences Is So Challenging: 2 / M.G. Claros et al. // Biology. Molecular Diversity Preservation International. – 2012. – Vol. 1, № 2. – P. 439-459.
7. McLoughlin, S. Fossil Plants: Gymnosperms / S. McLoughlin // Encyclopedia of Geology. – Elsevier, 2021. – P. 476-500.
8. Brenner, E.D. Using Genomics to Study Evolutionary Origins of Seeds / E.D. Brenner, D. Stevenson, ed. Williams Claire.G. // Landscapes, Genomics and Transgenic Conifers. – Dordrecht: Springer Netherlands. – 2006. – Vol. 9. – P. 85-106.
9. Sun, C. The Larix kaempferi genome reveals new insights into wood properties / C. Sun et al. // J Integr. Plant Biol. – 2022.
10. Niu, S. The Chinese pine genome and methylome unveil key features of conifer evolution / S. Niu et al. // Cell. – 2022. – Vol. 185, № 1. – P. 204-217.e14.
11. Mosca, E. A Reference Genome Sequence for the European Silver Fir (Abies alba Mill.): A Community-Generated Genomic Resource / E. Mosca et al. // G3 (Bethesda). – 2019. – Vol. 9, № 7. – P. 2039-2049.
12. Kuzmin, D.A. Stepwise large genome assembly approach: a case of Siberian larch (Larix sibirica Ledeb) / D.A. Kuzmin et al // BMC Bioinformatics. – 2019. – Vol. 20, № 1. – P. 37.

13. Zimin, A.V. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing / A.V. Zimin et al. // GigaScience. – 2017. – Vol. 6, № 1. – P. giw016.
14. Neale, D.B. The Douglas-Fir Genome Sequence Reveals Specialization of the Photosynthetic Apparatus in Pinaceae / D.B. Neale et al. // G3 Genes|Genomes|Genetics. – 2017. – Vol. 7, № 9. – P. 3157-3167.
15. Gonzalez-Ibeas, D. Assessing the Gene Content of the Megagenome: Sugar Pine (Pinus lambertiana) / D. Gonzalez-Ibeas et al. // G3 (Bethesda). – 2016. – Vol. 6, № 12. – P. 3787-3802.
16. Warren, R.L. Improved white spruce (Picea glauca) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism / R.L. Warren et al. // The Plant Journal. – 2015. – Vol. 83, № 2. – P. 189-212.
17. Nystedt, B. The Norway spruce genome sequence and conifer genome evolution / B. Nystedt et al. // Nature. – 2013. – Vol. 497, № 7451. – P. 579-584.
18. Semerikov, V.L. Nuclear and cytoplasmic variation within and between Eurasian Larix (Pinaceae) species / V.L. Semerikov, M. Lascoux // Am J Bot. – 2003. – Vol. 90, № 8. – P. 1113-1123.
19. Абаимов, В.Ф. Дендрология : учебное пособие для студентов высших учебных заведений, обучающихся по специальности «Лесное хозяйство» / В.Ф. Абаимов. – 3-е изд., перераб. – Москва: Академия, 2009. – 362 с.
20. Bondar, E.I. Annotation of Siberian Larch (Larix sibirica Ledeb.) Nuclear Genome – One of the Most Cold-Resistant Tree Species in the Only Deciduous GENUS in Pinaceae: 15 / E.I. Bondar et al. // Plants. – Multidisciplinary Digital Publishing Institute, 2022. – Vol. 11, № 15. – P. 2062.
21. Neale, D.B. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies / D.B. Neale et al. // Genome Biology. – 2014. – Vol. 15, № 3. – P. R59.
22. Zimin, A. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome / A. Zimin et al. // Genetics. – 2014. – Vol. 196, № 3. – P. 875-890.
23. Stevens, K. A. et al. Sequence of the Sugar Pine Megagenome / K. A. Stevens et al. // Genetics. – 2016. – P. 34.
24. Dulamsuren, C. Diverging climate trends in Mongolian taiga forests influence growth and regeneration of Larix sibirica / C. Dulamsuren et al. // Oecologia. – 2010. – Vol. 163, № 4. – P. 1091-1102.
25. Semerikov, V.L. Southern montane populations did not contribute to the recolonization of West Siberian Plain by Siberian larch (Larix sibirica): a range-

wide analysis of cytoplasmic markers: 19 / V.L. Semerikov et al. // Molecular Ecology. – 2013. – Vol. 22, № 19. – P. 4958-4971.

26. Tumenjargal, B. Physical and mechanical properties of wood and their geographic variations in Larix sibirica trees naturally grown in Mongolia: 1 / B. Tumenjargal et al. // Sci Rep. – Nature Publishing Group, 2020. – Vol. 10, № 1. – P. 12936.
27. Babushkina, E.A. The effect of individual genetic heterozygosity on general homeostasis, heterosis and resilience in Siberian larch (Larix sibirica Ledeb.) using dendrochronology and microsatellite loci genotyping / E.A. Babushkina et al. // Dendrochronologia. – 2016. – Vol. 38. – P. 26-37.
28. Oreshkova, N.V. Genetic diversity, population structure, and differentiation of Siberian larch, Gmelin larch and Cajander larch on SSR-markers data / N.V. Oreshkova, M.M. Belokon, S. Jamiyansuren // Russian Journal of Genetics. – 2013. – Vol. 49, № 2. – P. 178-186.
29. Oreshkova, N.V. Development of Microsatellite Genetic Markers in Siberian larch (Larix sibirica Ledeb.) Based on the De Novo Whole Genome Sequencing / N.V. Oreshkova et al. // Russian Journal of Genetics. – 2017.
30. Oreshkova, N.V. Development of Nuclear Microsatellite Markers with Long (Tri-, Tetra-, Penta-, and Hexanucleotide) Motifs for Three Larch Species Based on the de novo Whole Genome Sequencing of Siberian Larch (Larix sibirica Ledeb.) / N.V. Oreshkova et al. // Russ J Genet. – 2019. – Vol. 55, № 4. – P. 444-450.
31. Krutovsky, K.V. Postgenomic technologies in practical forestry: development of genome-wide markers for timber origin identification and other applications / K.V. Krutovsky et al. // Forestry Engineering Journal. – 2019. – Vol. 9, № 1. – P. 9-16.
32. Bondar, E.I. Siberian larch (Larix sibirica Ledeb.) chloroplast genome and development of polymorphic chloroplast markers / E.I. Bondar et al. // BMC Bioinformatics. – 2019. – Vol. 20, № 1. – P. 38.
33. Putintseva, Y.A. Siberian larch (Larix sibirica Ledeb.) mitochondrial genome assembled using both short and long nucleotide sequence reads is currently the largest known mitogenome / Y.A. Putintseva et al. // BMC Genomics. – 2020. – Vol. 21, № 1. – P. 654.
34. Abril, J.F. Genome Annotation // Encyclopedia of Bioinformatics and Computational Biology / J.F. Abril, S. Castellano, ed. S. Ranganathan et al. – Oxford: Academic Press, 2019. – P. 195-209.
35. Keddy, P. Evolutionary ecology of plant-plant interactions: an empirical modelling approach / P. Keddy // Annals of Botany. – 2007. – Vol. 99, № 2. – P. 372-374.
36. Gregory, T.R. Eukaryotic genome size databases / T.R. Gregory et al. // Nucleic Acids Res. – 2007. – Vol. 35, № Database issue. – P. D332-338.

37. Leitch, I.J. Genome sizes through the ages: 2 / I.J. Leitch // Heredity. – Nature Publishing Group, 2007. –Vol. 99, № 2. – P. 121-122.
38. Fleischmann, A. et al. Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms / A. Fleischmann et al. // Annals of Botany. – 2014. – Vol. 114, № 8. – P. 1651-1663.
39. Pellicer, J. The largest eukaryotic genome of them all? / J. Pellicer, M.F. Fay, I.J. Leitch // Botanical Journal of the Linnean Society. – 2010. – Vol. 164, № 1. – P. 10-15.
40. Sterck, L. How many genes are there in plants (… and why are they there)? / L. Sterck et al. // Current Opinion in Plant Biology. – 2007. – Vol. 10, № 2. – P. 199-203.
41. Cheng, C.-Y. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome / C.-Y. Cheng et al. // The Plant Journal. – John Wiley & Sons, Ltd, 2017. – Vol. 89, № 4. – P. 789-804.
42. Tang, H. An improved genome release (version Mt4.0) for the model legume Medicago truncatula / Tang H. et al. // BMC Genomics. – 2014. – Vol. 15, № 1. – P. 312.
43. Huang, G. Genome sequence of Gossypium herbaceum and genome updates of Gossypium arboreum and Gossypium hirsutum provide insights into cotton A-genome evolution: 5 / Huang G. et al. // Nat Genet. – Nature Publishing Group, 2020. – Vol. 52, № 5. – P. 516-524.
44. Neale, D.B. Assembled and annotated 26.5 Gbp coast redwood genome: a resource for estimating evolutionary adaptive potential and investigating hexaploid origin / D.B. Neale et al. // G3 (Bethesda). – 2022. – Vol. 12, № 1. – P. jkab380.
45. Schoch, C.L. NCBI Taxonomy: a comprehensive update on curation, resources and tools / C.L. Schoch et al. // Database (Oxford). – 2020. – Vol. 2020. – P. baaa062.
46. Brenner, E.D. Using Genomics to Study Evolutionary Origins of Seeds / E.D. Brenner, D. Stevenson, ed. G. Williams Claire // Landscapes, Genomics and Transgenic Conifers. – Dordrecht: Springer Netherlands, 2006. – Vol. 9. – P. 85-106.
47. Wan, T. A genome for gnetophytes and early evolution of seed plants: 2 / T. Wan et al. // Nature Plants. – Nature Publishing Group, 2018. – Vol. 4, № 2. – P. 82-89.
48. Berardini, T.Z. The arabidopsis information resource: Making and mining the «gold standard» annotated reference plant genome / T.Z. Berardini et al. // Genesis. – 2015. – Vol. 53, № 8. – P. 474-485.

49. Badouin, H. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution: 7656 / H. Badouin et al.// Nature. – Nature Publishing Group, 2017. – Vol. 546, № 7656. – P. 148-152.
50. Zimin, A.V. The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum / A.V. Zimin et al. // Gigascience. – 2017. – Vol. 6, № 11. – P. 1-7.
51. Wan, T. The Welwitschia genome reveals a unique biology underpinning extreme longevity in deserts: 1 / T. Wan et al. // Nat Commun. – Nature Publishing Group, 2021. – Vol. 12, № 1. – P. 4247.
52. Liu, Y. The Cycas genome and the early evolution of seed plants: 4 / Y. Liu et al. // Nat. Plants. – Nature Publishing Group, 2022. – Vol. 8, № 4. – P. 389-401.
53. Li, Z. Early genome duplications in conifers and other seed plants / Z. Li et al. // Sci Adv. – 2015. – Vol. 1, № 10. – P. e1501084.
54. Perera, D. Exploring the loblolly pine (Pinus taeda L.) genome by BAC sequencing and Cot analysis / D. Perera et al. // Gene. – 2018. – Vol. 663. – P. 165-177.
55. Wegrzyn, J.L. Unique Features of the Loblolly Pine (Pinus taeda L.) Megagenome Revealed Through Sequence Annotation / J.L. Wegrzyn et al. // Genetics. – 2014. – Vol. 196, № 3. – P. 891-909.
56. Lee, S.-I. Transposable Elements and Genome Size Variations in Plants / S.-I. Lee, N.-S. Kim // Genomics Inform. – Korea Genome Organization, 2014. – Vol. 12, № 3. – P. 87-97.
57. Губанов, И. А. Дикорастущие полезные растения СССР / И. А. Губанов, И.Л. Крылова, В. Л. Тихонова, под ред. Т. А. Работнова. – Москва: Мысль, 1976. – 360 с.
58. Лесная энциклопедия : в 2-х томах /гл.ред. Г. И. Воробьев и др. – Т. 2. –Москва: Советская энциклопедия, 1985. – 631 с.
59. Дылис, Н.В. Лиственница / Н.В. Дылис. – Москва: Лесная промышленность, 1981. – 97 с.
60. Лучник, З.И. Интродукция деревьев и кустарников в Алтайском крае / З.И. Лучник. –Москва: Колос, 1970. – 656 с.
61. Batalova, A.Y. Comparative Genomics of Seasonal Senescence in Forest Trees: 7 / A.Y. Batalova et al. // International Journal of Molecular Sciences. – Multidisciplinary Digital Publishing Institute, 2022. – Vol. 23, № 7. – P. 3761.
62. Кашин, В.И. Лиственничные леса Европейского севера России / В.И. Кашин, А.С. Козобродов. – Архангельск: АФРГО РАН, 1994. – 215 с.

63. Орешкова, Н.В. Разработка микросателлитных маркеров лиственницы сибирской (Larix sibirica Ledeb.) на основе полногеномного de novo секвенирования / Н.В. Орешкова и др. // Генетика. – 2017. – Vol. 53, № 11.
64. Salzberg, S.L. Next-generation genome annotation: we still struggle to get it right / S.L. Salzberg // Genome Biology. – 2019. – Vol. 20, № 1. – P. 92.
65. Pertea, M. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise / M. Pertea et al. // Genome Biol. – 2018. – Vol. 19, № 1. – P. 208.
66. Novák, P. et al. Repeat-sequence turnover shifts fundamentally in species with large genomes / P. Novák et al. // Nat Plants. – 2020. – Vol. 6, № 11. – P. 1325-1329.
67. Storer, J.M. Methodologies for the De novo Discovery of Transposable Element Families / J.M. Storer et al. // Genes (Basel). – 2022. – Vol. 13, № 4. – P. 709.
68. Chakraborty, M. Evolution of genome structure in the Drosophila simulans species complex / M.Chakraborty et al. // Genome Res. – 2021. – Vol. 31, № 3. – P. 380-396.
69. SanMiguel, P. Nested Retrotransposons in the Intergenic Regions of the Maize Genome / P. SanMiguel et al. // Science. – American Association for the Advancement of Science, 1996. – Vol. 274, № 5288. – P. 765-768.
70. Vargiu, L. Classification and characterization of human endogenous retroviruses; mosaic forms are common / L.Vargiu et al. // Retrovirology. – 2016. – Vol. 13, № 1. – P. 7.
71. Arkhipova, I.R. Giant Transposons in Eukaryotes: Is Bigger Better? / I.R. Arkhipova, I.A. Yushenova // Genome Biology and Evolution. – 2019. – Vol. 11, № 3. – P. 906–918.
72. Bao, Z. Automated de novo identification of repeat sequence families in sequenced genomes / Z. Bao, S.R. Eddy // Genome Res. – 2002. – Vol. 12, № 8. – P. 1269-1276.
73. Volfovsky, N. A clustering method for repeat analysis in DNA sequences / N. Volfovsky, B.J. Haas, S.L. Salzberg // Genome Biology. – 2001. – Vol. 2, № 8. – P. research0027.1.
74. Price, A.L. De novo identification of repeat families in large genomes / A.L. Price, N.C. Jones, P.A. Pevzner // Bioinformatics. – 2005. – Vol. 21 Suppl 1. – P. i351-358.
75. Sohrab V. TEfinder: A Bioinformatics Pipeline for Detecting New Transposable Element Insertion Events in Next-Generation Sequencing Data: 2 / V. Sohrab et al. // Genes. –Multidisciplinary Digital Publishing Institute, 2021. – Vol. 12, № 2. – P. 224.

76. Xu Z. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons / Z. Xu, H. Wang // Nucleic Acids Research. – 2007. – Vol. 35, № suppl_2. – P. W265–W268.
77. Steinbiss S. et al. Fine-grained annotation and classification of de novo predicted LTR retrotransposons // Nucleic Acids Research. 2009. Vol. 37, № 21. P. 7002–7013.
78. Ellinghaus, D. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons / D. Ellinghaus, S. Kurtz, U. Willhoeft // BMC Bioinformatics. – 2008. – Vol. 9, № 1. – P. 18.
79. Han, Y. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences / Y. Han, S. R. Wessler // Nucleic Acids Research. – 2010. – Vol. 38, № 22. – P. e199.
80. Sharp, P.A. The discovery of split genes and RNA splicing / P.A. Sharp // Trends in Biochemical Sciences. – 2005. – Vol. 30, № 6. – P. 279-281.
81. Frey, K. Animal, Fungi, and Plant Genome Sequences Harbor Different Non-Canonical Splice Sites / K. Frey, B. Pucker // Cells. – 2020. – Vol. 9, № 2. – P. 458.
82. Pucker, B. Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes / B. Pucker, S.F. Brockington // BMC Genomics. – 2018. – Vol. 19, № 1. – P. 980.
83. Plotkin, J.B. Synonymous but not the same: the causes and consequences of codon bias: 1 / J.B. Plotkin, G. Kudla // Nat Rev Genet. – Nature Publishing Group, 2011. – Vol. 12, № 1. – P. 32-42.
84. Lukashin, A.V. GeneMark.hmm: new solutions for gene finding / A.V. Lukashin, M. Borodovsky // Nucleic Acids Res. – 1998. – Vol. 26, № 4. – P. 1107-1115.
85. Delcher A.L. Improved microbial gene identification with GLIMMER / A.L. Delcher et al. // Nucleic Acids Res. – 1999. – Vol. 27, № 23. – P. 4636-4641.
86. Su, M. Small proteins: untapped area of potential biological importance / M. Su et al. // Front Genet. – 2013. – Vol. 4. – P. 286.
87. Trapnell, C. TopHat: discovering splice junctions with RNA-Seq / C. Trapnell, L. Pachter, S.L. Salzberg // Bioinformatics. – 2009. – Vol. 25, № 9. – P. 1105-1111.
88. Dobin, A. STAR: ultrafast universal RNA-seq aligner / A. Dobin et al. // Bioinformatics. – 2013. – Vol. 29, № 1. – P. 15-21.
89. Kim, D. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype / D. Kim et al. // Nat Biotechnol. – 2019. – Vol. 37, № 8. – P. 907-915.
90. Stoler, N. Sequencing error profiles of Illumina sequencing instruments / N. Stoler, A. Nekrutenko / N. Stoler // NAR Genom Bioinform. – 2021. – Vol. 3, № 1. – P. lqab019.

91. Ewing, B. Base-calling of automated sequencer traces using phred. II. Error probabilities / B. Ewing, P. Green // Genome Res. – 1998. – Vol. 8, № 3. – P. 186-194.
92. Raghavan, V. A simple guide to de novo transcriptome assembly and annotation / V. Raghavan et al. // Brief Bioinform. – 2022. – Vol. 23, № 2. – P. bbab563.
93. Grabherr, M.G. Full-length transcriptome assembly from RNA-Seq data without a reference genome: 7 / M.G. Grabherr et al. // Nat Biotechnol. – Nature Publishing Group, 2011. – Vol. 29, № 7. – P. 644-652.
94. Xie, Y. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads / Y. Xie et al. // Bioinformatics. – 2014. –Vol. 30, № 12. – P. 1660-1666.
95. Bushmanova, E. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data / E. Bushmanova et al. // Gigascience. – 2019. – Vol. 8, № 9. – P. giz100.
96. Seppey, M. BUSCO: Assessing Genome Assembly and Annotation Completeness / M. Seppey M., M. Manni, E.M. Zdobnov // Methods Mol Biol. – 2019. – Vol. 1962. P. 227-245.
97. Zdobnov, E.M. OrthoDB in 2020: evolutionary and functional annotations of orthologs E.M. Zdobnov et al. // Nucleic Acids Res. – 2020. – Vol. 49, № D1. – P. D389-D393.
98. Bolger, M.E. Plant genome and transcriptome annotations: from misconceptions to simple solutions / M.E Bolger., B. Arsova, B. Usadel // Brief Bioinform. – 2017. – Vol. 19, № 3. – P. 437-449.
99. Krishnakumar, V. Araport: the Arabidopsis information portal / V. Krishnakumar et al. // Nucleic Acids Res. – 2015. – Vol. 43, № Database issue. – P. D1003-1009.
100. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana // Nature. – 2000. – Vol. 408, № 6814. – P. 796-815.
101. Hou, X. A near-complete assembly of an Arabidopsis thaliana genome / X. Hou et al. // Molecular Plant. – Elsevier, 2022. – Vol. 15, № 8. – P. 1247-1250.
102. Ashburner, M. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium / M. Ashburner et al. // Nat Genet. – 2000. – Vol. 25, № 1. – P. 25-29.
103. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine // Nucleic Acids Res. – 2021. – Vol. 49, № D1. – P. D325–D334.
104. Moriya, Y. KAAS: an automatic genome annotation and pathway reconstruction server / Y. Moriya et al. // Nucleic Acids Res. – 2007. – Vol. 35, № Web Server issue. – P. W182-185.

105. Lohse, M. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data / M. Lohse et al. // Plant Cell Environ. – 2014. – Vol. 37, № 5. – P. 1250-1258.
106. Conesa, A. Blast2GO: A comprehensive suite for functional analysis in plant genomics / A. Conesa, S. Götz // Int J Plant Genomics. – 2008. – Vol. 2008. – P. 619832.
107. Götz, S. High-throughput functional annotation and data mining with the Blast2GO suite / S. Götz et al. // Nucleic Acids Res. – 2008. – Vol. 36, № 10. – P. 3420-3435.
108. Reyes, A. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues / A. Reyes, W. Huber // Nucleic Acids Research. –2018. – Vol. 46, № 2. – P. 582-592.
109. Tatarinova, T. NPEST: a nonparametric method and a database for transcription start site prediction / T. Tatarinova et al. // Quant Biol. – 2013. – Vol. 1, № 4. – P. 261-271.
110. Juven-Gershon, T. Regulation of gene expression via the core promoter and the basal transcriptional machinery / T. Juven-Gershon, J.T. Kadonaga // Developmental Biology. –2010. – Vol. 339, № 2. – P. 225-229.
111. Alexandrov, N.N. Features of Arabidopsis Genes and Genome Discovered using Full-length cDNAs / N.N. Alexandrov et al. // Plant Mol Biol. – 2006. – Vol. 60, № 1. – P. 69-85.
112. Alexandrov, N.N. et al. Insights into corn genes derived from large-scale cDNA sequencing / N.N. Alexandrov et al. // Plant Mol Biol. – 2009. – Vol. 69, № 1-2. – P. 179-194.
113. Tatarinova, T.V. Nucleotide diversity analysis highlights functionally important genomic regions / T.V. Tatarinova et al. // Sci Rep. – 2016. – Vol. 6, № 1. – P. 35730.
114. Triska, M. Nucleotide patterns aiding in prediction of eukaryotic promoters / M. Triska et al. // PLOS ONE. – Public Library of Science, 2017. – Vol. 12, № 11. – P. e0187243.
115. Troukhan, M. Genome-Wide Discovery of cis-Elements in Promoter Sequences Using Gene Expression / M. Troukhan et al. // OMICS: A Journal of Integrative Biology. – Mary Ann Liebert, Inc., publishers, 2009. – Vol. 13, № 2. – P. 139-151.
116. Roy, A.L. Core promoters in transcription: old problem, new insights / A.L. Roy, D.S. Singer // Trends in Biochemical Sciences. – Elsevier, 2015. – Vol. 40, № 3. – P. 165-171.

117. Sandelin, A. Mammalian RNA polymerase II core promoters: insights from genome-wide studies / A. Sandelin et al. // Nat Rev Genet. – 2007. – Vol. 8, № 6. – P. 424-436.
118. Franco-Zorrilla, J.M. Identification of plant transcription factor target sequences / J.M. Franco-Zorrilla, R. Solano // Biochimica et Biophysica Acta (BBA) – Gene Regulatory Mechanisms. – 2017. – Vol. 1860, № 1. – P. 21-30.
119. Morton, T. Paired-End Analysis of Transcription Start Sites in Arabidopsis Reveals Plant-Specific Promoter Signatures / T. Morton et al. // The Plant Cell. – 2014. – Vol. 26, № 7. – P. 2746-2760.
120. Ibraheem, O. In silico analysis of cis-acting regulatory elements in 5′ regulatory regions of sucrose transporter gene families in rice (Oryza sativa Japonica) and Arabidopsis thaliana / O. Ibraheem, C.E.J. Botha, G. Bradley // Computational Biology and Chemistry. – 2010. Vol. 34, № 5. – P. 268-283.
121. Duraisamy, G.S. Identification and characterization of promoters and cis-regulatory elements of genes involved in secondary metabolites production in hop (Humulus lupulus. L) / G.S. Duraisamy et al. // Comput Biol. Chem. – 2016. – Vol. 64. – P. 346-352.
122. Wong, D.C.J. et al. Genome-wide analysis of cis-regulatory element structure and discovery of motif-driven gene co-expression networks in grapevine / D.C.J. Wong et al. // DNA Research. – 2017. – Vol. 24, № 3. – P. 311-326.
123. Kumari, S. Genome-Wide Computational Prediction and Analysis of Core Promoter Elements across Plant Monocots and Dicots / S. Kumari, D. Ware // PLOS ONE. – Public Library of Science, 2013. – Vol. 8, № 10. – P. e79011.
124. Peters, J.P. DNA curvature and flexibility in vitro and in vivo / J.P. Peters, L.J. Maher // Q Rev Biophys. 2010. – Vol. 43, № 1. – P. 23-63.
125. Gan Y. A comparison study on feature selection of DNA structural properties for promoter prediction / Y. Gan, J. Guan, S. Zhou // BMC Bioinformatics. – 2012. – Vol. 13, № 1. – P. 4.
126. Kanhere A. Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes / A. Kanhere, M. Bansal // Nucleic Acids Research. – 2005. – Vol. 33, № 10. – P. 3165-3175.
127. Triska, M. Analysis of cis-Regulatory Elements in Gene Co-expression Networks in Cancer / M.Triska et al. // Methods Mol Biol. – 2017. – Vol. 1613. – P. 291-310.
128. Yella, V.R. Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy / V.R. Yella, A. Kumar, M. Bansal // Sci Rep. – 2018. – Vol. 8, № 1. – P. 4520.

129. Kozobay-Avraham L. Curvature distribution in prokaryotic genomes / L. Kozobay-Avraham, S. Hosid, A. Bolshoy // In Silico Biol. – 2004. – Vol. 4, № 3. – P. 361-375.
130. Kumar, A. Unveiling DNA structural features of promoters associated with various types of TSSs in prokaryotic transcriptomes and their role in gene expression / A. Kumar, M. Bansal // DNA Res. – Oxford Academic, 2017. – Vol. 24, № 1. – P. 25-35.
131. Pandey, S.P. Computational analysis of plant RNA Pol-II promoters / S.P. Pandey, A. Krishnamachari // Biosystems. – 2006. – Vol. 83, № 1. – P. 38-50.
132. Zuo, Y.-C. Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-Skew and DNA geometric flexibility / Y.-C. Zuo, Q.-Z. Li // Genomics. – 2011. – Vol. 97, № 2. – P. 112-120.
133. Aerts, S. Comprehensive analysis of the base composition around the transcription start site in Metazoa / S. Aerts et al. // BMC Genomics. – 2004. – Vol. 5, № 1. – P. 34.
134. Fujimori, S. GC-compositional strand bias around transcription start sites in plants and fungi / S. Fujimori, T. Washio, M. Tomita // BMC Genomics. – 2005. – Vol. 6, № 1. – P. 26.
135. Tatarinova, T. Skew in CG content near the transcription start site in Arabidopsis thaliana / T. Tatarinova et al. // Bioinformatics. – 2003. – Vol. 19, № suppl_1. – P. i313-i314.
136. Carels, N.Two classes of genes in plants / Carels N., Bernardi G. // Genetics. – 2000. – Vol. 154, № 4. – P. 1819-1825.
137. Tatarinova, T.V. GC3 biology in corn, rice, sorghum and other grasses / T.V. Tatarinova et al. // BMC Genomics. – 2010. – Vol. 11, № 1. – P. 308.
138. Chan, K.-L. Evidence-based gene models for structural and functional annotations of the oil palm genome / K.-L. Chan et al. // Biology Direct. – 2017. – Vol. 12, № 1. – P. 21.
139. Clément, Y. The bimodal distribution of genic GC content is ancestral to monocot species / Y. Clément et al. // Genome Biol Evol. – 2014. – Vol. 7, № 1. – P. 336-348.
140. Serres-Giardi, L. Patterns and Evolution of Nucleotide Landscapes in Seed Plants / L. Serres-Giardi et al. // The Plant Cell. – 2012. – Vol. 24, № 4. – P. 1379-1397.
141. Хлесткина, Е.К. Молекулярные маркеры в генетических исследованиях и в селекции: 4/2 / Е.К. Хлесткина // Вавиловский журнал генетики и селекции. – 2013. – Vol. 17, №4/2. – P. 1044-1054.

142. Qin, Z. Evolution Analysis of Simple Sequence Repeats in Plant Genome / Z. Qin et al. // PLoS One. – 2015. – Vol. 10, № 12. – P. e0144108.
143. Калько, Г.В. ДНК-маркеры для оценки генетических ресурсов ели и сосны / Г.В. Калько // Труды Санкт-Петербургского Научно-Исследовательского Института Лесного Хозяйства. – 2015. – № 4.
144. Tereba, A. Analysis of DNA profiles of ash (Fraxinus excelsior L.) to provide evidence of illegal logging / A. Tereba et al. // Wood Sci Technol. – 2017. – Vol. 51, № 6. – P. 1377-1387.
145. Breidenbach, N. Genetic structure of coast redwood (Sequoia sempervirens [D. Don] Endl.) populations in and outside of the natural distribution range based on nuclear and chloroplast microsatellite markers / N. Breidenbach, O. Gailing, K.V. Krutovsky // PLOS ONE. – Public Library of Science, 2020. – Vol. 15, № 12. – P. e0243556.
146. Wang, X.-R. Molecular Markers in Population Genetics of Forest Trees / X.-R. Wang, A. Szmidt // Scandinavian Journal of Forest Research. – 2001. – Vol. 16. – P. 199-220.
147. Van Oosterhout, C. micro-checker: software for identifying and correcting genotyping errors in microsatellite data / C. Van Oosterhout et al. // Molecular Ecology Notes. – 2004. – Vol. 4, № 3. – P. 535-538.
148. Wheeler, G.L. A review of the prevalence, utility, and caveats of using chloroplast simple sequence repeats for studies of plant biology1 / G.L. Wheeler et al. // Appl Plant Sci. –2014. – Vol. 2, № 12. – P. apps.1400059.
149. Breidenbach, N. Development of novel polymorphic nuclear and chloroplast microsatellite markers in coast redwood (Sequoia sempervirens) / N. Breidenbach, O. Gailing, K.V. Krutovsky // Plant Genetic Resources. – Cambridge University Press, 2019. – Vol. 17, № 3. – P. 293-297.
150. Захаров-Гезехус, И.А. Цитоплазматическая наследственность: 1 / И.А. Захаров-Гезехус // Вавиловский журнал генетики и селекции. – 2014. – Vol. 18, № 1. – P. 93-102.
151. Hipkins, V. Organelle genomes in conifers: structure, evolution, and diversity / V. Hipkins, K. Krutovskii, S.H. Straws. – 1995.
152. Chen, X.-B. Development and characterization of polymorphic genic-SSR markers in Larix kaempferi / X.-B. Chen, Y.-H. Xie, X.-M. Sun // Molecules. – 2015. – Vol. 20, № 4. – P. 6060-6067.
153. Wagner, S. Two highly informative dinucleotide SSR multiplexes for the conifer Larix decidua (European larch) / S. Wagner, S. Gerber, R.J. Petit // Molecular Ecology Resources. – 2012. – Vol. 12, № 4. – P. 717-725.

154. Chen, C. Development and characterization of microsatellite loci in western larch (Larix occidentalis Nutt.) / C. Chen et al. // Mol Ecol Resour. – 2009. – Vol. 9, № 3. – P. 843-845.
155. Khasa, P.D. Isolation, characterization, and inheritance of microsatellite loci in alpine larch and western larch / P.D. Khasa et al. // Genome. – 2000. – Vol. 43, № 3. – P. 439-448.
156. Polezhaeva, M.A. Cytoplasmic DNA variation and biogeography of Larix Mill. in northeast Asia / M.A. Polezhaeva, M. Lascoux, V.L. Semerikov // Mol Ecol. – 2010. – Vol. 19, № 6. – P. 1239-1252.
157. Семериков, В.Л. Нуклеотидное разнообразие и неравновесие по сцеплению потенциально адаптивно-значимых генов Larix sibirica / В.Л. Семериков, С.А. Семерикова, М.А. Полежаева // Генетика. 2013. Vol. 49, № 9.
158. Holmen, J. et al. Cross-species amplification of 36 cyprinid microsatellite loci in Phoxinus phoxinus (L.) and Scardinius erythrophthalmus (L.) / J. Holmen et al. // BMC Research Notes. 2009. – Vol. 2, № 1. – P. 248.
159. Benson, G. Tandem repeats finder: a program to analyze DNA sequences / G. Benson // Nucleic Acids Res. – 1999. – Vol. 27, № 2. – P. 573-580.
160. Thiel, T. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.) / T. Thiel et al. // Theor Appl. Genet. – 2003. – Vol. 106, № 3. – P. 411-422.
161. Wang, X. GMATo: A novel tool for the identification and analysis of microsatellites in large genomes / X. Wang, P. Lu, Z. Luo // Bioinformation. – 2013. – Vol. 9, № 10. – P. 541-544.
162. Jewell, E. SSRPrimer and SSR Taxonomy Tree: Biome SSR discovery / E. Jewell et al. // Nucleic Acids Res. – 2006. – Vol. 34, № Web Server issue. – P. W656-659.
163. Asif, M. et al. High resolution MetaPhor agarose gel electrophoresis for genotyping with microsatellite markers / M. Asif et al. // Pakistan Journal of Agricultural Sciences. – 2008. –Vol. 45, № 1. – P. 75-79.
164. Smit, A. RepeatModeler Open-1.0 / A. Smit, R. Hubley. – 2008.
165. Smit, A. RepeatMasker Open-4.0 / A. Smit, R. Hubley, P. Green. – 2013.
166. Bao, W. Repbase Update, a database of repetitive elements in eukaryotic genomes / W. Bao, K.K. Kojima, O. Kohany // Mobile DNA. – 2015. – Vol. 6, № 1. – P. 11.
167. Abrusán, G. TEclass – a tool for automated classification of unknown eukaryotic transposable elements / G. Abrusán et al. // Bioinformatics. – 2009. – Vol. 25, № 10. – P. 1329-1330.

168. Nussbaumer, T. MIPS PlantsDB: a database framework for comparative plant genome research / T. Nussbaumer et al. // Nucleic Acids Res. – 2013. – Vol. 41, № Database issue. – P. D1144-D1151.
169. Wegrzyn, J.L. Insights into the Loblolly Pine Genome: Characterization of BAC and Fosmid Sequences / J.L. Wegrzyn et al. // PLOS ONE. – Public Library of Science, 2013. – Vol. 8, № 9. – P. e72439.
170. Kojima, K.K. Human transposable elements in Repbase: genomic footprints from fish to humans / K.K. Kojima // Mobile DNA. – 2018. – Vol. 9, № 1. – P. 2.
171. Mascagni, F. A comparison of methods for LTR-retrotransposon insertion time profiling in the Populus trichocarpa genome / F. Mascagni et al. // Caryologia. – Taylor & Francis, 2018. – Vol. 71, № 1. – P. 85-92.
172. Ou, S. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons1[OPEN] / S. Ou, N. Jiang // Plant Physiol. –2018. – Vol. 176, № 2. – P. 1410-1422.
173. Zhou, S.-S. A comprehensive annotation dataset of intact LTR retrotransposons of 300 plant genomes: 1 / S.-S. Zhou et al. // Sci Data. – Nature Publishing Group, 2021. – Vol. 8, № 1. – P. 174.
174. De La Torre, A.R. Contrasting Rates of Molecular Evolution and Patterns of Selection among Gymnosperms and Flowering Plants / A.R. De La Torre et al. // Molecular Biology and Evolution. – 2017. – Vol. 34, № 6. – P. 1363-1377.
175. Mistry, J. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions / J. Mistry et al. // Nucleic Acids Res. – 2013. – Vol. 41, № 12. – P. e121.
176. El-Gebali, S. The Pfam protein families database in 2019 / S. El-Gebali et al. // Nucleic Acids Res. – 2019. – Vol. 47, № D1. – P. D427-D432.
177. Holt, C. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects / C. Holt, M. Yandell // BMC Bioinformatics. – 2011. – Vol. 12, № 1. – P. 491.
178. Korf, I. Gene finding in novel genomes / I. Korf // BMC Bioinformatics. – 2004. – Vol. 5, № 1. – P. 59.
179. Lomsadze, A. et al. Gene identification in novel eukaryotic genomes by self-training algorithm / A. Lomsadze et al. // Nucleic Acids Research. – 2005. – Vol. 33, № 20. – P. 6494-6506.
180. Stanke, M. AUGUSTUS: ab initio prediction of alternative transcripts / M. Stanke et al. // Nucleic Acids Res. – 2006. – Vol. 34, № Web Server issue. – P. W435-439.

181. Scalzitti, N. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms / N. Scalzitti et al. // BMC Genomics. – 2020. – Vol. 21, № 1. – P. 293.
182. Trapnell, C. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks / C. Trapnell et al. // Nat Protoc. – 2012. – Vol. 7, № 3. – P. 562-578.
183. Pearson, W.R. An introduction to sequence similarity («homology») searching / W.R. Pearson W.R. // Curr Protoc Bioinformatics. – 2013. – Chapter 3, Unit3.1.
184. Jones, P. InterProScan 5: genome-scale protein function classification / P. Jones et al. // Bioinformatics. – 2014. – Vol. 30, № 9. – P. 1236-1240.
185. Blum, M. The InterPro protein families and domains database: 20 years on / M. Blum et al. // Nucleic Acids Research. – 2021. – Vol. 49, № D1. – P. D344-D354.
186. Benjamini, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing / Y. Benjamini, Y. Hochberg // Journal of the Royal Statistical Society. Series B (Methodological). – [Royal Statistical Society, Wiley], 1995. – Vol. 57, № 1. – P. 289-300.
187. Storey, J.D. A direct approach to false discovery rates / J.D. Storey // Journal of the Royal Statistical Society: Series B (Statistical Methodology). – 2002. – Vol. 64, № 3. – P. 479-498.
188. Wu, C.-S. Comparative Chloroplast Genomes of Pinaceae: Insights into the Mechanism of Diversified Genomic Organizations / C.-S. Wu et al. // Genome Biology and Evolution. –2011. – Vol. 3. – P. 309-319.
189. Parks, M. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes / M. Parks, R. Cronn, A. Liston // BMC Biology. – 2009. – Vol. 7, № 1. – P. 84.
190. Langmead, B. Fast gapped-read alignment with Bowtie 2 / B. Langmead, S.L. Salzberg // Nat Methods. – 2012. – Vol. 9, № 4. – P. 357-359.
191. Bankevich, A. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing / A. Bankevich et al. // Journal of Computational Biology. – 2012. – Vol. 19, № 5. – P. 455-477.
192. Boetzer, M. Scaffolding pre-assembled contigs using SSPACE / M. Boetzer et al. // Bioinformatics. – 2011. – Vol. 27, № 4. – P. 578-579.
193. Overbeek, R. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST) / R. Overbeek et al. // Nucl. Acids Res. – 2014. – Vol. 42, № D1. – P. D206-D214.
194. Andrews, S. FastQC: a quality control tool for high throughput sequence data / S. Andrews [Electronic resource]. – 2010. – URL:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed: 19.08.2022).

195. Bolger, A.M. Trimmomatic: a flexible trimmer for Illumina sequence data / A.M. Bolger, M. Lohse, B. Usadel // Bioinformatics. – 2014. – Vol. 30, № 15. – P. 2114-2120.
196. Sahlin, K. BESST--efficient scaffolding of large fragmented assemblies / K.Sahlin et al. // BMC Bioinformatics. – 2014. – Vol. 15. – P. 281.
197. Zimin, A.V. The MaSuRCA genome assembler / A.V. Zimin et al. // Bioinformatics. – 2013. – Vol. 29, № 21. – P. 2669-2677.
198. Hunt, M. REAPR: a universal tool for genome assembly evaluation / M. Hunt et al. // Genome Biol. – 2013. – Vol. 14, № 5. – P. R47.
199. Alverson, A.J. et al. Insights into the Evolution of Mitochondrial Genome Size from Complete Sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae) / A.J. Alverson et al. // Mol Biol Evol. – 2010. – Vol. 27, № 6. – P. 1436-1448.
200. Laslett, D. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences / D. Laslett, B. Canback // Nucleic Acids Res. – 2004. – Vol. 32, № 1. – P. 11-16.
201. Lowe, T.M. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence / T.M. Lowe, S.R. Eddy // Nucleic Acids Res. – 1997. – Vol. 25, № 5. – P. 955-964.
202. Lagesen, K. RNAmmer: consistent and rapid annotation of ribosomal RNA genes / K. Lagesen et al. // Nucleic Acids Res. – 2007. – Vol. 35, № 9. – P. 3100-3108.
203. Birol, I. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data / I. Birol et al. // Bioinformatics. – 2013. – Vol. 29, № 12. – P. 1492-1497.
204. Bondar, E.I. Genome-Wide Prediction of Transcription Start Sites in Conifers: 3 / E.I. Bondar et al. // International Journal of Molecular Sciences. – Multidisciplinary Digital Publishing Institute, 2022. – Vol. 23, № 3. – P. 1735.
205. Shahmuradov, I.A. TSSPlant: a new tool for prediction of plant Pol II promoters / I.A. Shahmuradov, R.Kh. Umarov, V.V. Solovyev // Nucleic Acids Research. – 2017. – Vol. 45, № 8. – P. e65.
206. Rangannan, V. High-quality annotation of promoter regions for 913 bacterial genomes / V. Rangannan, M. Bansal // Bioinformatics. – 2010. – Vol. 26, № 24. – P. 3043-3050.
207. Kel, A.E. MATCH: A tool for searching transcription factor binding sites in DNA sequences / A.E. Kel et al. // Nucleic Acids Res. – 2003. – Vol. 31, № 13. – P. 3576-3579.

208. Devey, M.E. A genetic linkage map for Pinus radiata based on RFLP, RAPD, and microsatellite markers / M.E. Devey et al. // Theoret. Appl. Genetics. – 1996. – Vol. 92, № 6. – P. 673-679.
209. Коропачинский, И.Ю. Древесные растения Азиатской России / И.Ю. Коропачинский, Т.Н. Встовская. – Новосибирск: Академическое издательство «ГЕО», 2012. – 707 с.
210. Абаимов, А.П. Лиственницы Гмелина и Каяндера / А.П. Абаимов, И.Ю. Коропачинский. – Новосибирск: Наука, 1984. – 120 с.
211. Guries, R.P. Genetic Diversity and Population Structure in Pitch Pine (Pinus rigida Mill.) / R.P. Guries, F.T. Ledig // Evolution. [Society for the Study of Evolution, Wiley], 1982. – Vol. 36, № 2. – P. 387-402.
212. Nei, M. Genetic Distance between Populations / M. Nei // The American Naturalist. [University of Chicago Press, American Society of Naturalists], 1972. – Vol. 106, № 949. – P. 283-292.
213. Peakall, R. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research –an update / R. Peakall, P.E. Smouse // Bioinformatics. – 2012. – Vol. 28, № 19. – P. 2537-2539.
214. Belyayev, A. Bursts of transposable elements as an evolutionary driving force / A. Belyayev // Journal of Evolutionary Biology. – 2014. – Vol. 27, № 12. – P. 2573-2584.
215. Naville, M. Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements / M. Naville et al. // Current Biology. – 2019. – Vol. 29, № 7. – P. 1161-1168.e6.
216. Piegu, B. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice / B. Piegu et al. // Genome Res. – 2006. – Vol. 16, № 10. – P. 1262-1269.
217. Tsukahara, S. Bursts of retrotransposition reproduced in Arabidopsis: 7262 / S. Tsukahara et al. // Nature. Nature Publishing Group, 2009. – Vol. 461, № 7262. – P. 423-426.
218. Zeh, D.W. Transposable elements and an epigenetic basis for punctuated equilibria Zeh D.W., Zeh J.A., Ishida Y. // BioEssays. – 2009. – Vol. 31, № 7. – P. 715-726.
219. Kelly, L.J. et al. Analysis of the giant genomes of Fritillaria (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size: 2 / L.J. Kelly et al. // New Phytol. – 2015. – Vol. 208, № 2. – P. 596-607.
220. Wang, W. Astonishing 35S rDNA diversity in the gymnosperm species Cycas revoluta Thunb / W. Wang et al. // Chromosoma. – 2016. – Vol. 125, № 4. – P. 683-699.

221. Magbanua, Z.V. Adventures in the Enormous: A 1.8 Million Clone BAC Library for the 21.7 Gb Genome of Loblolly Pine / Z.V. Magbanua et al. // PLOS ONE. – Public Library of Science, 2011. – Vol. 6, № 1. – P. e16214.
222. Titievsky, A. Comparative Genomics Analysis of Repetitive Elements in Ten Gymnosperm Species: «Dark Repeatome» and Its Abundance in Conifer and Gnetum Species / A. Titievsky et al. // Life (Basel). – 2021. – Vol. 11, № 11. – P. 1234.
223. Civáň, P.On the Coevolution of Transposable Elements and Plant Genomes / P. Civáň, M. Švec, P. Hauptvogel // Journal of Botany. – Hindawi, 2011. – Vol. 2011. – P. e893546.
224. Arkhipova, I.R. Distribution and Phylogeny of Penelope-Like Elements in Eukaryotes: 6 I.R. Arkhipova // Systematic Biology. – 2006. – Vol. 55, № 6. – P. 875-885.
225. Evgen'ev, M.B. Penelope-like elements – a new class of retroelements: distribution, function and possible evolutionary significance / M.B. Evgen'ev, I.R. Arkhipova // CGR. – Karger Publishers, 2005. – Vol. 110, № 1-4. – P. 510-521.
226. Lin X. An Ancient Transkingdom Horizontal Transfer of Penelope-Like Retroelements from Arthropods to Conifers / X. Lin, N. Faridi, C. Casola // Genome Biol Evol. – 2016. –Vol. 8, № 4. – P. 1252-1266.
227. Gao, D. Horizontal Transfer of Non-LTR Retrotransposons from Arthropods to Flowering Plants Gao D. et al. // Molecular Biology and Evolution. – 2018. – Vol. 35, № 2. – P. 354-364.
228. Heitkam, T. Comparative Repeat Profiling of Two Closely Related Conifers (Larix decidua and Larix kaempferi) Reveals High Genome Similarity With Only Few Fast-Evolving Satellite DNAs // Frontiers in Genetics. – 2021. – Vol. 12.
229. Wicker, T. A unified classification system for eukaryotic transposable elements / T. Wicker et al. // Nat Rev Genet. – 2007. – Vol. 8, № 12. – P. 973-982.
230. Zhang, L. The structure and retrotransposition mechanism of LTR-retrotransposons in the asexual yeast Candida albicans: 6 / L. Zhang et al. // Virulence. – 2014. – Vol. 5, № 6. – P. 655-664.
231. Aroh, O. Genome-wide characterization of LTR retrotransposons in the non-model deep-sea annelid Lamellibrachia luymesi / O. Aroh, K.M. Halanych // BMC Genomics. – 2021. –Vol. 22, № 1. – P. 466.
232. Barghini, E. Identification and characterisation of Short Interspersed Nuclear Elements in the olive tree (Olea europaea L.) genome / E. Barghini et al. // Mol Genet Genomics. – 2017. – Vol. 292, № 1. – P. 53-61.

233. Kumar, A. Plant Retrotransposons / A. Kumar, J.L. Bennetzen // Annual Review of Genetics. – 1999. – Vol. 33, № 1. – P. 479-532.
234. Brunner, S. Evolution of DNA Sequence Nonhomologies among Maize Inbreds / S. Brunner et al. // The Plant Cell. – 2005. – Vol. 17, № 2. – P. 343-360.
235. Paterson, A.H. The Sorghum bicolor genome and the diversification of grasses: 7229 / A.H. Paterson et al. // Nature. – Nature Publishing Group, 2009. – Vol. 457, № 7229. – P. 551-556.
236. Buti, M. Temporal dynamics in the evolution of the sunflower genome as revealed by sequencing and annotation of three large genomic regions / M. Buti et al. // Theor Appl. Genet. – 2011. – Vol. 123, № 5. – P. 779.
237. Zhao, M.-L. Induction of jasmonate signalling regulators MaMYC2s and their physical interactions with MaICE1 in methyl jasmonate-induced chilling tolerance in banana fruit: 1 / M.-L. Zhao et al. // Plant, Cell & Environment. – 2013. – Vol. 36, № 1. – P. 30-51.
238. Yin, H. Genome-wide Annotation and Comparative Analysis of Long Terminal Repeat Retrotransposons between Pear Species of P. bretschneideri and P. Communis: 1 / H. Yin et al. // Sci Rep. – Nature Publishing Group, 2015. – Vol. 5, № 1. – P. 17644.
239. Jones, J.D.G. The plant immune system: 7117 / J.D.G. Jones, J.L. Dangl // Nature. Nature Publishing Group, 2006. – Vol. 444, № 7117. – P. 323-329.
240. Kobe, B. The leucine-rich repeat as a protein recognition motif / B. Kobe, A.V. Kajava // Curr Opin Struct Biol. – 2001. – Vol. 11, № 6. – P. 725-732.
241. Schaper, E. The evolution and function of protein tandem repeats in plants / E. Schaper, M. Anisimova // New Phytol. – 2015. – Vol. 206, № 1. – P. 397-410.
242. Song, H. Evolutionary balance between LRR domain loss and young NBS–LRR genes production governs disease resistance in Arachis hypogaea cv. Tifrunner: 1 / H. Song et al. // BMC Genomics. – 2019. – Vol. 20, № 1. – P. 844.
243. Koralewski,T.E. Evolution of Exon-Intron Structure and Alternative Splicing / T.E. Koralewski, K.V. Krutovsky // PLOS ONE. – Public Library of Science, 2011. – Vol. 6, № 3. – P. e18055.
244. Pellicer, J. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies: 2 / J. Pellicer, I.J. Leitch // New Phytologist. – 2020. – Vol. 226, № 2. – P. 301-305.
245. Eilbeck, K. The Sequence Ontology: a tool for the unification of genome annotations / K. Eilbeck et al. // Genome Biology. – 2005. – Vol. 6, № 5. – P. R44.
246. Eilbeck, K. Quantitative measures for the management and comparison of annotated genomes / K. Eilbeck et al. // BMC Bioinformatics. – 2009. – Vol. 10, № 1. – P. 67.

247. Sena, J.S. Evolution of gene structure in the conifer Picea glauca: a comparative analysis of the impact of intron size / J.S. Sena et al. // BMC Plant Biol. – 2014. – Vol. 14. – P. 95.
248. Niklas, K.J. The Cell Walls that Bind the Tree of Life / K.J. Niklas // BioScience. – 2004. – Vol. 54, № 9. – P. 831-841.
249. Sarkar, P. Plant cell walls throughout evolution: towards a molecular understanding of their design principles / P. Sarkar, E. Bosneaga, M. Auer // Journal of Experimental Botany. – 2009. – Vol. 60, № 13. – P. 3615-3635.
250. Li, L. The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding sinapyl alcohol dehydrogenase / L. Li et al. // Plant Cell. – 2001. – Vol. 13, № 7. – P. 1567-1586.
251. Hatfield, R. Lignin Formation in Plants. The Dilemma of Linkage Specificity / R. Hatfield, W. Vermerris // Plant Physiology. – 2001. – Vol. 126, № 4. – P. 1351-1357.
252. Wagner, A. Chapter 2 – Lignification and Lignin Manipulations in Conifers / A. Wagner, L. Donaldson, J. Ralph // Advances in Botanical Research / ed. L. Jouanin, C. Lapierre. – Academic Press, 2012. – Vol. 61. – P. 37-76.
253. Pascual, M.B. Biosynthesis and Metabolic Fate of Phenylalanine in Conifers / M.B. Pascual et al. // Front Plant Sci. – 2016. – Vol. 7. – P. 1030.
254. Porth, I. Defense mechanisms against herbivory in Picea: sequence evolution and expression regulation of gene family members in the phenylpropanoid pathway / I. Porth et al. // BMC Genomics. – 2011. – Vol. 12, № 1. – P. 608.
255. Yadav, V. Phenylpropanoid Pathway Engineering: An Emerging Approach towards Plant Defense / V. Yadav et al. // Pathogens. – 2020. – Vol. 9, № 4. – P. 312.
256. Vogt, T. Phenylpropanoid Biosynthesis: 1 / T. Vogt // Molecular Plant. – 2010. – Vol. 3, № 1. – P. 2-20.
257. Bagal, U.R. The phenylalanine ammonia lyase (PAL) gene family shows a gymnosperm-specific lineage / U.R. Bagal et al. // BMC Genomics. – 2012. – Vol. 13, № 3. – P. S1.
258. El-Azaz, J. Identification of a small protein domain present in all plant lineages that confers high prephenate dehydratase activity / J. El-Azaz et al. // Plant J. – 2016. – Vol. 87, № 2. – P. 215-229.
259. van Doorn, W.G. Morphological classification of plant cell deaths: 8 / W.G. van Doorn et al. // Cell Death Differ. – Nature Publishing Group, 2011. – Vol. 18, № 8. – P. 1241-1246.

260. Klim, J. Ancestral State Reconstruction of the Apoptosis Machinery in the Common Ancestor of Eukaryotes / J. Klim et al. // G3 Genes|Genomes|Genetics. – 2018. – Vol. 8, № 6. – P. 2121-2134.
261. Hara-Nishimura, I. The role of vacuole in plant cell death: 8 / I. Hara-Nishimura, N. Hatsugai // Cell Death Differ. – Nature Publishing Group, 2011. – Vol. 18, № 8. – P. 1298-1304.
262. Minina, E.A. Vacuolar cell death in plants: Metacaspase releases the brakes on autophagy / E.A. Minina, A.P. Smertenko, P.V. Bozhkov // Autophagy. – 2014. – Vol. 10, № 5. – P. 928-929.
263. Reape, T.J. Programmed cell death in plants: distinguishing between different modes / T.J. Reape, E.M. Molony, P.F. McCabe // Journal of Experimental Botany. – 2008. – Vol. 59, № 3. – P. 435-444.
264. van Doorn, W.G. Classes of programmed cell death in plants, compared to those in animals: 14 / W.G. van Doorn // Journal of Experimental Botany. – 2011. – Vol. 62, № 14. – P. 4749-4761.
265. Kalra, G. Senescence and Programmed Cell Death // Plant Physiology, Development and Metabolism / G. Kalra, S.C. Bhatla, ed. S.C. Bhatla, A. Lal M. – Singapore: Springer, 2018. – P. 937-966.
266. Delorme, V.G. A matrix metalloproteinase gene is expressed at the boundary of senescence and programmed cell death in cucumber / V.G. Delorme et al. // Plant Physiol. – 2000. – Vol. 123, № 3. – P. 917-927.
267. Valandro, F. Programmed cell death (PCD) control in plants: New insights from the Arabidopsis thaliana deathosome / F. Valandro et al. // Plant Science. – 2020. – Vol. 299. – P. 110603.
268. Van Hautegem, T. Only in dying, life: programmed cell death during plant development: 2 / T. Van Hautegem et al. // Trends Plant Sci. – 2015. – Vol. 20, № 2. – P. 102-113.
269. Koyama, T. The roles of ethylene and transcription factors in the regulation of onset of leaf senescence / T. Koyama // Frontiers in Plant Science. – 2014. – Vol. 5.
270. Kim, S.-H. Genes for Plant Autophagy: Functions and Interactions / S.-H. Kim et al. // Mol Cells. – 2012. – Vol. 34, № 5. – P. 413-423.
271. Daneva, A. Functions and Regulation of Programmed Cell Death in Plant Development / A. Daneva et al. // Annu Rev Cell Dev Biol. – 2016. – Vol. 32. – P. 441-468.
272. Zhu, J.-K. Salt and drought stress signal transduction in plants / J.-K. Zhu // Annu Rev Plant Biol. – 2002. – Vol. 53. – P. 247-273.

273. Xue-Xuan, X. Biotechnological implications from abscisic acid (ABA) roles in cold stress and leaf senescence as an important signal for improving plant sustainable survival under abiotic-stressed conditions / X. Xue-Xuan et al. // Crit Rev Biotechnol. – 2010. – Vol. 30, № 3. – P. 222-230.
274. Preston, J. Adaptation to seasonality and the winter freeze / J. Preston, S. Sandve // Frontiers in Plant Science. – 2013. – Vol. 4.
275. Wu, J. Linkage of cold acclimation and disease resistance through plant–pathogen interaction pathway in Vitis amurensis grapevine / J. Wu et al. // Funct Integr Genomics. – 2014. – Vol. 14, № 4. – P. 741-755.
276. Costa-Broseta, Á. Nitric Oxide Controls Constitutive Freezing Tolerance in Arabidopsis by Attenuating the Levels of Osmoprotectants, Stress-Related Hormones and Anthocyanins: 1 / Á. Costa-Broseta et al. // Sci Rep. – Nature Publishing Group, 2018. – Vol. 8, № 1. – P. 9268.
277. Song, Y. Abscisic Acid as an Internal Integrator of Multiple Physiological Processes Modulates Leaf Senescence Onset in Arabidopsis thaliana / Y. Song et al. // Frontiers in Plant Science. – 2016. – Vol. 7.
278. Xu, P. Transcription factor CDF4 promotes leaf senescence and floral organ abscission by regulating abscisic acid and reactive oxygen species pathways in Arabidopsis / P. Xu, H. Chen, W. Cai // EMBO Rep. – 2020. – Vol. 21, № 7. – P. e48967.
279. Raab, S. Identification of a novel E3 ubiquitin ligase that is required for suppression of premature senescence in Arabidopsis / S. Raab et al.// Plant J. – 2009. – Vol. 59, № 1. – P. 39-51.
280. Lee, I.C. Age-Dependent Action of an ABA-Inducible Receptor Kinase, RPK1, as a Positive Regulator of Senescence in Arabidopsis Leaves / I.C. Lee et al. // Plant and Cell Physiology. – 2011. – Vol. 52, № 4. – P. 651-662.
281. Breeze, E. High-resolution temporal profiling of transcripts during Arabidopsis leaf senescence reveals a distinct chronology of processes and regulation / E. Breeze et al. // Plant Cell. – 2011. – Vol. 23, № 3. – P. 873-894.
282. Yang J. A NAP-AAO3 Regulatory Module Promotes Chlorophyll Degradation via ABA Biosynthesis in Arabidopsis Leaves / J. Yang, E. Worley, M. Udvardi // Plant Cell. – 2014. – Vol. 26, № 12. – P. 4862-4874.
283. Groen, S.C. The evolution of ethylene signaling in plant chemical ecology / S.C. Groen, N.K. Whiteman // J Chem Ecol. – 2014. – Vol. 40, № 7. – P. 700-716.
284. Wang, C. Insights into the Origin and Evolution of the Plant Hormone Signaling Machinery / C. Wang et al. // Plant Physiology. – 2015. – Vol. 167, № 3. – P. 872-886.

285. Han, G.-Z. Evolution of jasmonate biosynthesis and signaling mechanisms / G.-Z. Han // Journal of Experimental Botany. – 2017. – Vol. 68, № 6. – P. 1323-1331.
286. Truman, W. Arabidopsis systemic immunity uses conserved defense signaling pathways and is mediated by jasmonates: 3 / W. Truman et al. // Proc Natl Acad Sci USA. – 2007. – Vol. 104, № 3. – P. 1075-1080.
287. Ali, M.S., Baek, K.-H. Jasmonic Acid Signaling Pathway in Response to Abiotic Stresses in Plants / M.S. Ali, K.-H. Baek // Int J Mol Sci. – 2020. – Vol. 21, № 2. – P. E621.
288. Qiu, Z. Exogenous jasmonic acid can enhance tolerance of wheat seedlings to salt stress / Z.Qiu et al. // Ecotoxicol Environ Saf. – 2014. – Vol. 104. – P. 202-208.
289. Todaka, D. Recent advances in the dissection of drought-stress regulatory networks and strategies for development of drought-tolerant transgenic rice plants / D. Todaka, K. Shinozaki, K. Yamaguchi-Shinozaki // Front Plant Sci. – 2015. – Vol. 6. – P. 84.
290. Mohamed, H.I. Improvement of drought tolerance of soybean plants by using methyl jasmonate / H.I. Mohamed, H.H. Latif // Physiol Mol Biol Plants. – 2017. – Vol. 23, – № 3. P. 545-556.
291. Fan, L. Amelioration of postharvest chilling injury in cowpea (Vigna sinensis) by methyl jasmonate (MeJA) treatments / L. Fan et al. // Scientia Horticulturae. – 2016. – Vol. 203. – P. 95-101.
292. Cerrudo, I. Low red/far-red ratios reduce Arabidopsis resistance to Botrytis cinerea and jasmonate responses via a COI1-JAZ10-dependent, salicylic acid-independent mechanism / I. Cerrudo et al. // Plant Physiol. – 2012. – Vol. 158, № 4. – P. 2042-2052.
293. Svyatyna, K. Light-dependent regulation of the jasmonate pathway / K. Svyatyna, M. Riemann // Protoplasma. – 2012. – Vol. 249 Suppl. 2. – P. S137-145.
294. Mewis, I. UV-B Irradiation Changes Specifically the Secondary Metabolite Profile in Broccoli Sprouts: Induced Signaling Overlaps with Defense Response to Biotic Stressors I.Mewis et al. // Plant Cell Physiol. – 2012. – Vol. 53, № 9. – P. 1546-1560.
295. Kozlowski, G. Methyl jasmonate protects Norway spruce [Picea abies (L.) Karst.] seedlings against Pythium ultimum Trow. / G. Kozlowski, A. Buchala, J.-P. Métraux // Physiological and Molecular Plant Pathology. – 1999. – Vol. 55, № 1. – P. 53-58.
296. Franceschi, V.R. Application of methyl jasmonate on Picea abies (Pinaceae) stems induces defense-related responses in phloem and xylem / V.R. Franceschi, T. Krekling, E. Christiansen // Am J Bot. – 2002. – Vol. 89, № 4. – P. 578-586.

297. Szmidt, A.E. Paternal inheritance of chloroplast DNA in Larix / A.E. Szmidt, T. Aldén, J.-E. Hällgren // Plant Mol Biol. – 1987. – Vol. 9, № 1. – P. 59-64.
298. Hipkins, V.D. Organelle genome in conifers: structure, evolution / V.D. Hipkins, K.V. Krutovskii, S.H. Straws // Forest Genetics. – 1994. – Vol. 1, № 4. – P. 179-189.
299. Whittall, J.B. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines / J.B. Whittall et al. // Molecular Ecology. – 2010. – Vol. 19, № s1. – P. 100-114.
300. Cronn, R. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology / R. Cronn et al. // Nucleic Acids Research. – 2008. – Vol. 36, № 19. – P. e122.
301. Willyard, A. Fossil Calibration of Molecular Divergence Infers a Moderate Mutation Rate and Recent Radiations for Pinus / A. Willyard et al. // Molecular Biology and Evolution. –2007. – Vol. 24, № 1. – P. 90-101.
302. Gernandt, D.S. Use of Simultaneous Analyses to Guide Fossil-Based Calibrations of Pinaceae Phylogeny / D.S. Gernandt et al. // International Journal of Plant Sciences. – The University of Chicago Press, 2008. – Vol. 169, № 8. – P. 1086-1099.
303. Lidholm, J. A three-step model for the rearrangement of the chloroplast trnK -psbA region of the gymnosperm Pinus contorts / J. Lidholm, P. Gustafsson // Nucleic Acids Research. –1991. – Vol. 19, № 11. – P. 2881-2887.
304. Provan, J. A low mutation rate for chloroplast microsatellites / J. Provan et al. // Genetics. – 1999. – Vol. 153, № 2. – P. 943-947.
305. Ebert, D. Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species / D. Ebert, R. Peakall // Molecular Ecology Resources. – 2009. – Vol. 9, № 3. – P. 673-690.
306. Afzal-Rafii Z. Chloroplast DNA supports a hypothesis of glacial refugia over postglacial recolonization in disjunct populations of black pine (Pinus nigra) in western Europe / Z. Afzal-Rafii, R.S. Dodd // Molecular Ecology. – 2007. – Vol. 16, № 4. – P. 723-736.
307. Höhn, M. Variation in the chloroplast DNA of Swiss stone pine (Pinus cembra L.) reflects contrasting post-glacial history of populations from the Carpathians and the Alps / M. Höhn et al. // Journal of Biogeography. – 2009. – Vol. 36, № 9. – P. 1798-1806.
308. Moreno-Letelier, A. Phylogeographic structure of Pinus strobiformis Engelm. across the Chihuahuan Desert filter-barrier / A. Moreno-Letelier, D. Piñero // Journal of Biogeography. – 2009. – Vol. 36, № 1. – P. 121-131.

309. Morse, A.M. Evolution of genome size and complexity in Pinus / A.M. Morse et al. // PLoS One. – 2009. – Vol. 4, № 2. – P. e4332.
310. Pritham, E.J. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses / E.J. Pritham, T. Putliwala, C. Feschotte // Gene. –2007. – Vol. 390, № 1-2. – P. 3-17.
311. Haapa-Paananen, S. Phylogenetic analysis of Maverick/Polinton giant transposons across organisms / S. Haapa-Paananen S., N. Wahlberg, H. Savilahti // Mol Phylogenet Evol. – 2014. – Vol. 78. – P. 271-274.
312. Handa, H. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (Brassica napus L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana / H. Handa // Nucleic Acids Res. – 2003. – Vol. 31, № 20. – P. 5907-5916.
313. Kim, B. Completion of the mitochondrial genome sequence of onion (Allium cepa L.) containing the CMS-S male-sterile cytoplasm and identification of an independent event of the ccmF N gene split / B. Kim et al. // Curr Genet. – 2016. – Vol. 62, № 4. – P. 873-885.
314. Rice, D.W. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella / D.W. Rice et al. // Science. – 2013. – Vol. 342, № 6165. – P. 1468-1473.
315. Sloan, D.B. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates / D.B. Sloan et al. // PLoS Biol. – 2012. – Vol. 10, № 1. – P. e1001241.
316. Boore, J.L. Animal mitochondrial genomes / Boore J.L. // Nucleic Acids Res. – 1999. – Vol. 27, № 8. – P. 1767-1780.
317. Mower, J.P. Plant Mitochondrial Genome Diversity: The Genomics Revolution / J.P. Mower, D.B. Sloan, A.J. Alverson // Plant Genome Diversity Volume 1: Plant Genomes, their Residents, and their Evolutionary Dynamics / ed. Wendel J.F. et al. – Vienna: Springer, 2012. – P. 123-144.
318. Gualberto, J.M. Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation / J.M. Gualberto, K.J. Newton // Annu Rev Plant Biol. – 2017. – Vol. 68. P. 225-252.
319. Chevigny, N. DNA Repair and the Stability of the Plant Mitochondrial Genome / N. Chevigny et al. // Int J Mol Sci. – 2020. – Vol. 21, № 1. – P. 328.
320. Warren, J.M. Linear Plasmids and the Rate of Sequence Evolution in Plant Mitochondrial Genomes / J.M. Warren et al. // Genome Biol Evol. – 2016. – Vol. 8, № 2. – P. 364-374.

321. Kan, S.-L. The complete mitochondrial genome of Taxus cuspidata (Taxaceae): eight protein-coding genes have transferred to the nuclear genome / S.-L. Kan et al. // BMC Evolutionary Biology. – 2020. – Vol. 20, № 1. – P. 10.
322. Kozik, A. et al. The alternative reality of plant mitochondrial DNA: One ring does not rule them all / A. Kozik et al. // PLoS Genet. – 2019. – Vol. 15, № 8. – P. e1008373.
323. Liberatore, K.L. The role of mitochondria in plant development and stress tolerance / K.L.Liberatore et al. // Free Radic Biol Med. – 2016. – Vol. 100. – P. 238-256.
324. Alverson, A.J. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber / A.J. Alverson et al. // Plant Cell. – 2011. – Vol. 23, № 7. – P. 2499-2513.
325. Wang, D. Plastid sequences contribute to some plant mitochondrial genes / D. Wang, M. Rousseau-Gueutin, J.N. Timmis // Mol Biol Evol. – 2012. – Vol. 29, № 7. – P. 1707-1711.
326. Gandini, C.L. Foreign Plastid Sequences in Plant Mitochondria are Frequently Acquired Via Mitochondrion-to-Mitochondrion Horizontal Transfer / C.L. Gandini, M.V. Sanchez-Puerta // Sci Rep. – 2017. – Vol. 7. – P. 43402.
327. Wolfe, K.H. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs / K.H. Wolfe, W.H. Li, P.M. Sharp // Proc Natl Acad Sci USA. – 1987. – Vol. 84, № 24. – P. 9054–9058.
328. Smith, D.R. Mutation rates in plastid genomes: they are lower than you might think / D.R. Smith // Genome Biol Evol. – 2015. – Vol. 7, № 5. – P. 1227-1234.
329. Lynch, M. Mutation pressure and the evolution of organelle genomic architecture / M. Lynch, B. Koskella, S. Schaack // Science. – 2006. – Vol. 311, № 5768. – P. 1727-1730.
330. Zhu, A. Unprecedented heterogeneity in the synonymous substitution rate within a plant genome / A. Zhu et al. // Mol Biol Evol. – 2014. – Vol. 31, № 5. – P. 1228-1236.
331. Sloan, D.B. One ring to rule them all? Genome sequencing provides new insights into the “master circle” model of plant mitochondrial DNA structure / D.B. Sloan // New Phytol. – 2013. – Vol. 200, № 4. – P. 978-985.
332. Biłas, R. et al. Cis-regulatory elements used to control gene expression in plants / R. Biłas et al. // Plant Cell Tiss Organ Cult. – 2016. – Vol. 127, № 2. – P. 269-287.
333. Messing, J. Plant Gene Structure // Genetic Engineering of Plants: An Agricultural Perspective / J. Messing et al., ed. Kosuge T. et al. – Boston, MA: Springer US, 1983. – P. 211-227.

334. Porto, M.S. Plant Promoters: An Approach of Structure and Function / M.S. Porto et al. // Mol Biotechnol. – 2014. – Vol. 56, № 1. – P. 38-49.
335. Dhadi, S.R. Genome-wide comparative analysis of putative bidirectional promoters from rice, Arabidopsis and Populus / S.R. Dhadi, N. Krom, W. Ramakrishna // Gene. – 2009. – Vol. 429, № 1-2. – P. 65-73.
336. Krom, N. Comparative Analysis of Divergent and Convergent Gene Pairs and Their Expression Patterns in Rice, Arabidopsis, and Populus / N. Krom, W. Ramakrishna // Plant Physiol. – Oxford Academic, 2008. – Vol. 147, № 4. – P. 1763-1773.
337. Yamamoto, Y.Y. et al. Characteristics of Core Promoter Types with respect to Gene Structure and Expression in Arabidopsis thaliana / Y.Y. Yamamoto et al. // DNA Research. – 2011. – Vol. 18, № 5. – P. 333-342.
338. Tian, F. et al. PlantRegMap: charting functional regulatory maps in plants / F. Tian et al. // Nucleic Acids Research. – 2020. – Vol. 48, № D1. – P. D1104-D1113.
339. Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation / E. Wingender // Brief Bioinform. – 2008. – Vol. 9, № 4. – P. 326-332.
340. Wasserman, W.W. Applied bioinformatics for the identification of regulatory elements / W.W. Wasserman, A. Sandelin // Nat Rev Genet. – 2004. – Vol. 5, № 4. – P. 276-287.
341. Dietz, K.-J. AP2/EREBP transcription factors are part of gene regulatory networks and integrate metabolic, hormonal and environmental signals in stress acclimation and retrograde signalling / K.-J. Dietz, M.O. Vogel, A. Viehhauser // Protoplasma. – 2010. – Vol. 245, № 1-4. – P. 3-14.
342. Liu, C. Expansion and stress responses of the AP2/EREBP superfamily in cotton / Liu C., Zhang T. // BMC Genomics. – 2017. – Vol. 18, № 1. – P. 118.
343. Noyes, M.B. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites / M.B. Noyes et al. // Cell. – 2008. – Vol. 133, № 7. – P. 1277-1289.
344. Svingen, T. Hox transcription factors and their elusive mammalian gene targets / T. Svingen, K.F. Tonissen // Heredity (Edinb). – 2006. – Vol. 97, № 2. – P. 88-96.
345. Guo, M. The Plant Heat Stress Transcription Factors (HSFs): Structure, Regulation, and Function in Response to Abiotic Stresses / M. Guo et al. // Front Plant Sci. – 2016. – Vol. 7. – P. 114.
346. Miller, G. Could heat shock transcription factors function as hydrogen peroxide sensors in plants? / G. Miller, R. Mittler // Ann Bot. – 2006. – Vol. 98, № 2. – P. 279-288.

347. Prouse, M.B. Interactions between the R2R3-MYB transcription factor, AtMYB61, and target DNA binding sites / M.B. Prouse, M.M. Campbell // PLoS One. – 2013. – Vol. 8, № 5. – P. e65132.
348. Wang, B. Structural insights into target DNA recognition by R2R3-MYB transcription factors / B. Wang et al. // Nucleic Acids Res. – 2020. – Vol. 48, № 1. – P. 460-471.
349. Choi, K. Arabidopsis meiotic crossover hotspots overlap with H2A.Z nucleosomes at gene promoters / K. Choi et al. // Nat Genet. – 2013. – Vol. 45, № 11. – P. 10.1038/ng.2766.
350. Hellsten, U. Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing / U. Hellsten et al. // Proc Natl Acad Sci USA. – 2013. – Vol. 110, № 48. – P. 19478-19482.
351. Glémin, S. GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis / S. Glémin et al. // Trends Genet. – 2014. – Vol. 30, № 7. – P. 263-270.
352. Wong, G.K.-S. Compositional Gradients in Gramineae Genes / G.K.-S. Wong et al. // Genome Res. – 2002. – Vol. 12, № 6. – P. 851-856.
353. Fortes, G.G. Diversity in isochore structure among cold-blooded vertebrates based on GC content of coding and non-coding sequences / G.G. Fortes et al. // Genetica. – 2007. – Vol. 129, № 3. – P. 281-289.
354. Jørgensen, F.G. Heterogeneity in Regional GC Content and Differential Usage of Codons and Amino Acids in GC-Poor and GC-Rich Regions of the Genome of Apis mellifera / F.G. Jørgensen, M.H. Schierup, A.G. Clark // Molecular Biology and Evolution. – 2007. – Vol. 24, № 2. – P. 611-619.
355. Lynch, D.B. Chromosomal G + C Content Evolution in Yeasts: Systematic Interspecies Differences, and GC-Poor Troughs at Centromeres / D.B. Lynch et al.// Genome Biology and Evolution. – 2010. – Vol. 2. – P. 572-583.
356. Ларионова А.Я., Орешкова (Яхнева) Н.В., Абаимов А.П. Генетическое разнообразие и дифференциация популяции лиственницы Гмелина в Эвенкии (Средняя Сибирь): 2–3 / А.Я. Ларионова, Н.В. Орешкова (Яхнева), А.П. Абаимов // Генетика. Россия, Красноярск: Федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева», 2004. – Vol. 40, № 9. – P. 1370-1377.
357. Knowles P. et al. Significant levels of self-fertilization in natural populations of tamarack / Knowles P. et al. // Canadian journal of botany. – 1987. – Vol. 65, № 6. – P. 1087-1091.

358. Oreshkova N.V. et al. Genetic diversity, structure and differentiation of Gmelin larch (Larix gmelinii (Rupr.) Rupr.) populations from Central Evenkia and Eastern Zabaikalje / N.V. Oreshkova et al. // Eurasian J. Forest Res. – 2006. – Vol. 9, № 1. – P. 1-8.
359. Яхнева (Орешкова), Н.В. Генетико-таксономический анализ популяций лиственницы Гмелина (Larix gmelinii (Rupr.) Rupr.) : дисс. … канд. биол. наук : 03.00.05 / Н.В. Яхнева (Орешкова):– Красноярск, 2004. – 157 с.
360. Орешкова, Н.В. Генетическая и фенотипическая изменчивость лиственницы Каяндера (Larix cajanderi Mayr.) на севере российского Дальнего Востока / Н.В. Орешкова, В.П. Ветрова, Н.В. Синельникова // Сибирский Экологический Журнал. – 2015. – Vol. 22, № 1. – P. 13-27.

ПРИЛОЖЕНИЯ

Приложение А

Относительное содержание семейств и классов повторов в сборке генома лиственницы сибирской, аннотированных с помощью комбинированной библиотеки повторов.

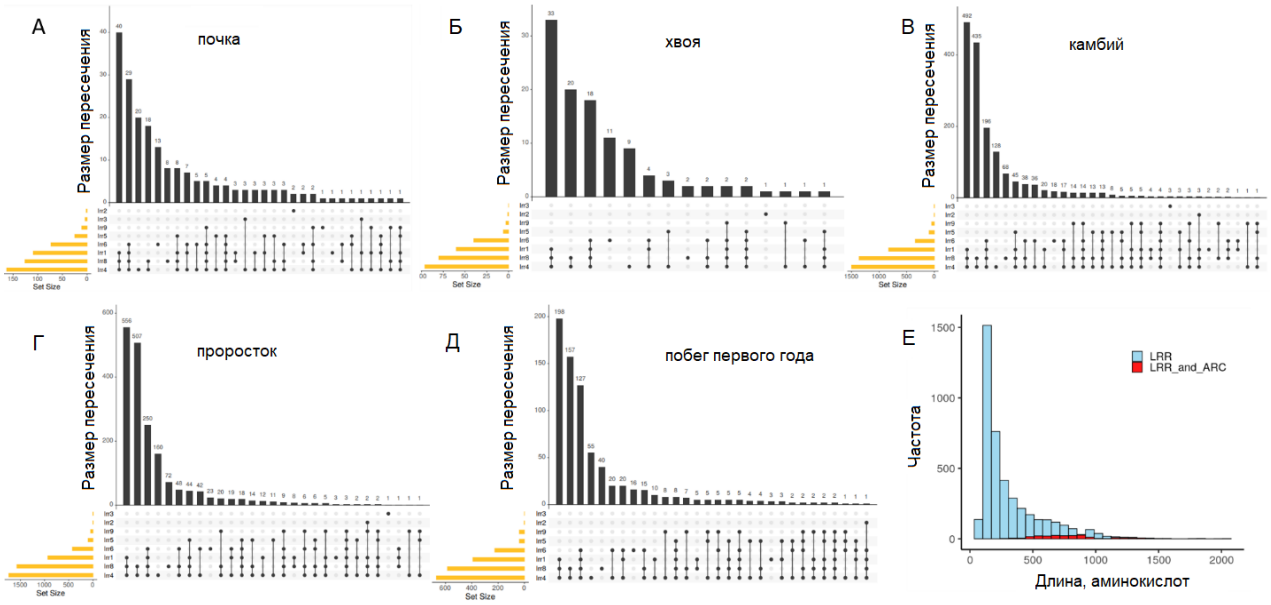
Суперсемейство	**Длина, п.н.**	**Количество**	**Доля от длины генома, %**
Class I: LTR retrotransposon			
Gypsy	192527487	878785	1,60
Gypsy/PtTalladega	1663158	3646	0,01
Gypsy/IFG	254098	990	<0,01
Gypsy/PtAppalachian	208844	773	
Gypsy/PtOuachita	191626	1025	
Gypsy/Gymny	138565	476	
Gypsy/PtBastrop	94602	594	
Gypsy/RLG	62107	1004	
Gypsy/PtOzark	24665	234	
Gypsy/ALISEI	21741	209	
Gypsy/PsAppalachian	16203	108	
Gypsy/PtAngelina	497	4	
Total *Gypsy*	**195203593**	**887848**	**1,63**
Copia	125043599	627489	1,04
Copia/PtConagree	115681	731	**<0,01**
Copia/Silava_Pta	115133	581	
Copia/PtPineywoods	64007	401	
Copia/PtCumberland	32521	251	
Total *Copia*	**125370941**	**629453**	**1,04**
DIR	7936310	51295	0,07
ERV	444525	7717	<0.01
BEL	418006	7389	
DIRS	135235	2287	
ATGP	22493	272	
RNLTR	10324	197	
RMER	6732	110	
RIRE	6682	100	
Unclassified LTR	1703439143	6246018	14,19
Total LTR	**2032993984**	**7832686**	**16,94**
Class I: Non-LTR retrotransposon			
LINE	1124723825	4237879	9,37
LINE/I	265525048	1577725	2,21
LINE/L1	21091550	94925	0,18
Total *LINE*	**1411340423**	**5910529**	**11,76**
Penelope	15970080	79330	0,13
Penelope/Poseidon	39219	411	<0,01
Total *Penelope*	**16009299**	**79741**	**0,13**

Продолжение Таблицы А

Суперсемейство	Длина, п.н.	Количество	Доля от длины генома, %
SINE	10566091	79177	0,09
LOA	19191	275	<0,01
Jockey	15690	120	
TRAS	14525	228	
LIN#_SM	10491	169	
BovB	9128	155	
Outcast	6887	121	
Hero	6599	122	
Other non-LTR	22649056	133018	0,19
Total non-LTR	**1460647380**	**6203655**	**12,17**
Other retrotransposons	390376007	1771180	3,25
Total Class I	**3884017371**	**15807521**	**32,36**
Class II: DNA transposon			
TIR	19582568	105413	0,16
Helitron	6739134	28439	0,06
EnSpm	2142128	28885	0,02
MuDR	587040	8191	<0,01
MuDR/Vandal	29783	546	
MuDR/Rehavkus	15474	220	
MuDR/Arnold	13638	151	
MuDR/ATMU	8652	143	
MuDR/OSMU	6609	85	
Total *MuDR*	**661196**	**9336**	**<0,01**
Mariner	462692	7270	<0,01
REP	167449	1355	
Maverick	517499	5211	
hAT	1073811	15821	
Mite	32189	505	
TcMar	30674	446	
Harbinger	20160	316	
Stowaway	3957	72	
PiggyBac	1065	20	
Tourist	845	16	
Unclassified	539184594	2226370	4,49
Total Class II	**570619961**	**2429475**	**4,76**
Others			
Unclassified	262494312	1345055	2,19
Simple repeats	46499317	983566	0,39
Host	12847092	18499	0,11
Low complexity	11560642	211886	0,09
Other	1215941	76131	0.01
tRNA	299376	4982	<0,01
rRNA	189617	1947	
RNA	1325	16	
Caulimovirus	401990	3767	<0,01
Grand Total	**4790146944**	**20882845**	**39,91**

Приложение Б

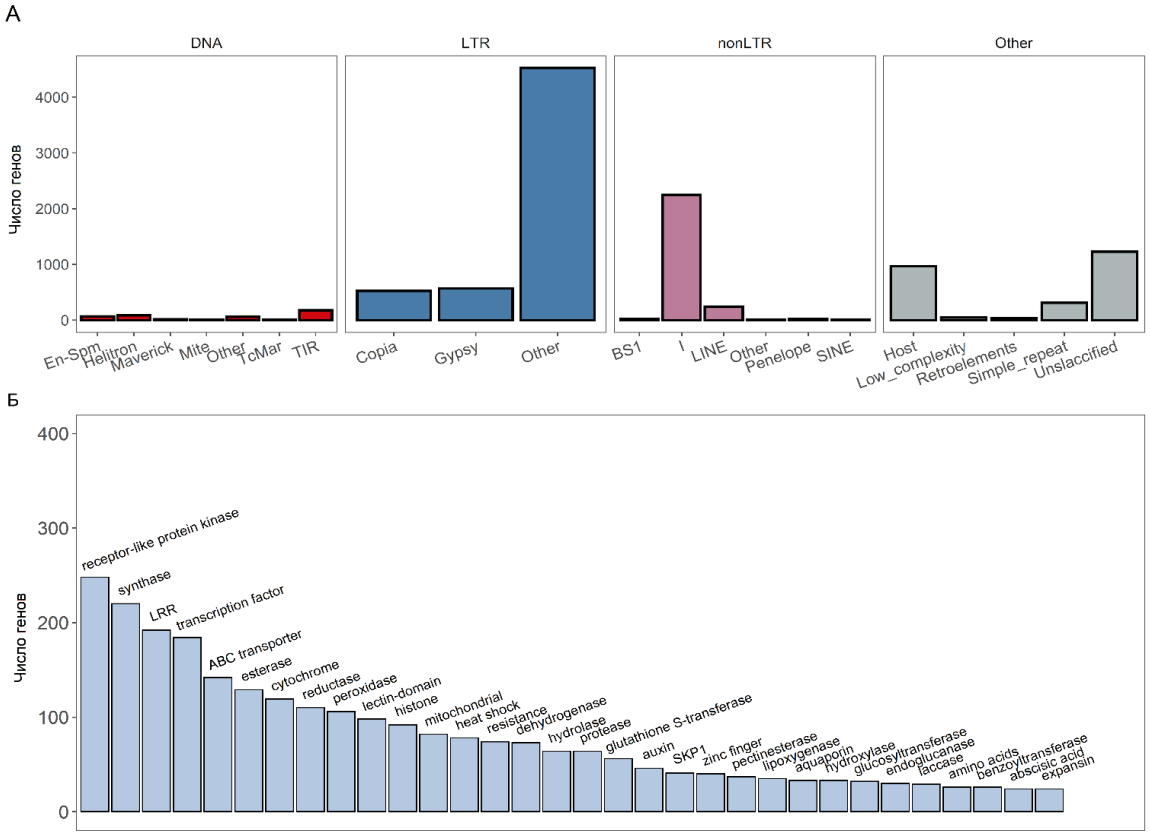
LRR гены



А–Д — количество транскриптов содержащих LRRs найденных с использованием девяти семейств (lrr1-lrr9) в транскриптомах пяти тканей (например, в ткани почки 43 транскрипта найдены в каждом из семейств LRR-1, LRR-4 и LRR-8), Е — распределение длин аминокислотных последовательностей предполагаемых R-генов, LRR — транскрипты, содержащие LRR domain (голубой), LRR и ARC — транскрипты, содержащие LRR и ARC домены (красный).

Приложение В

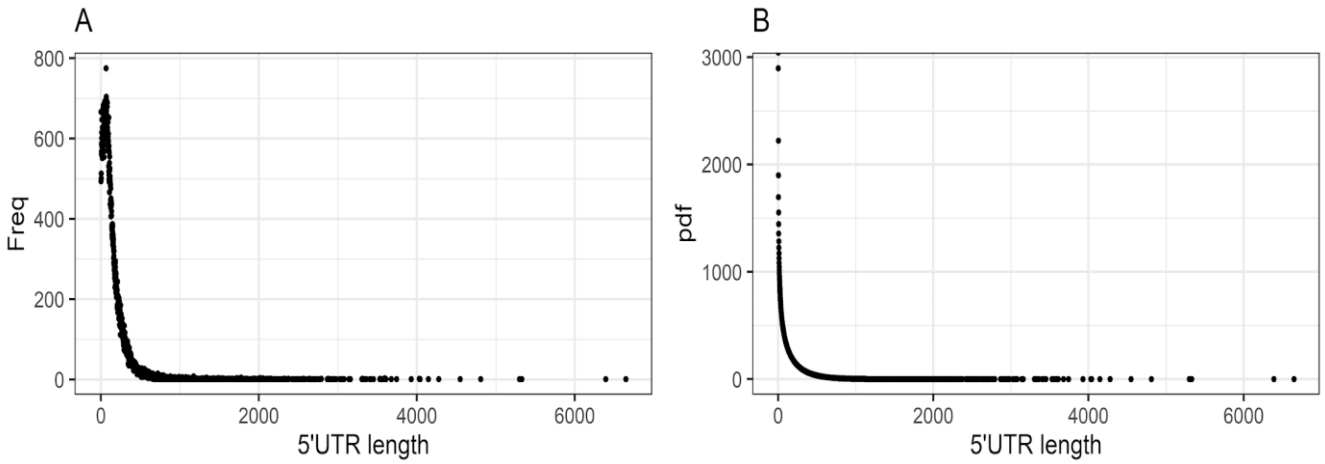
Повторы, пересекающиеся с генными моделями



А — число повторов, пересекающихся с кодирующей частью гена по крайней мере на 20%; Б — наиболее частые функциональные категории генов, имеющих по крайней мере 20% пересечение с повторами.

Приложение Г.1

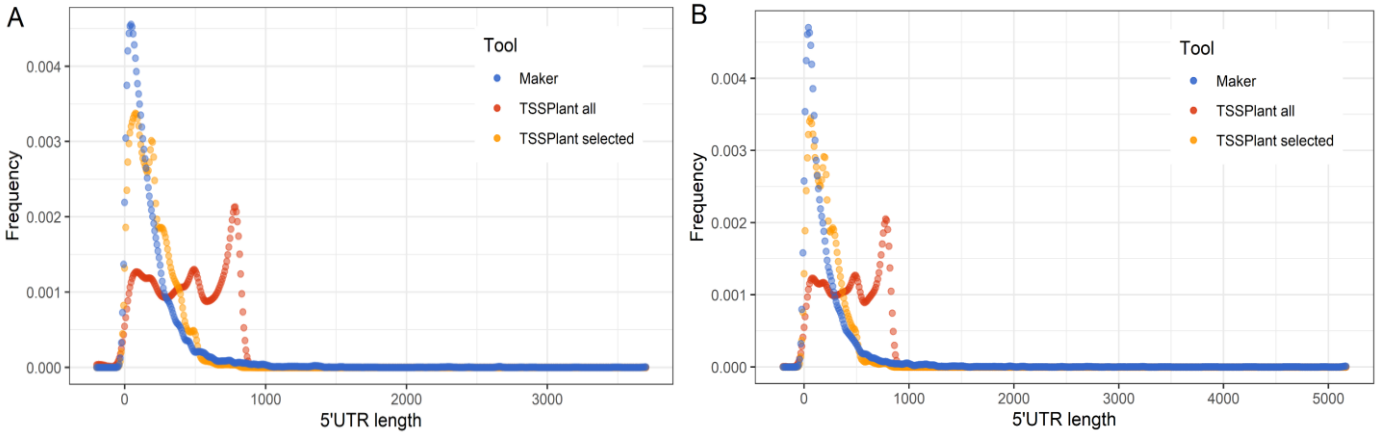
Распределение длин 5’UTR на основе генных аннотаций *A. thaliana*, *P. trichocarpa*, *O. sativa* и *S. bicolor*



А — частота встречаемости, Б — функция плотность вероятности распределения длин 5’UTR.

Приложение Г.2

Сравнение распределения длин of 5’UTR предсказанных MAKER и TSSPlant



А — в геноме *L. sibirica*, В — в геноме *P. glauca*

Приложение Д.1

60 локусов, отобранных для первичного тестирования

Локус	Мотив	Последовательности праймеров	Размеры продукта
Ls_1247092(2)	$CTT_{(19)}$	FW: AAACGCCAACCACAGAATTTAC RV: TTATCAAACGAGGGATGCTTCT	214
Ls_3956537	$TCA_{(19)}$	FW: TAGCAAGGAATTGTGGGATCA RV: GGATGATGATGCTGATGAAGATAG	261
Ls_611965	$CAG_{(18)}$	FW: GCAAGTTCCGATTCAAGAGTTT RV: GTTGCTGCTGAGGCATGTAG	236
Ls_3542003	$TCA_{(18)}$	FW: AACGATTTGGCATTTGACTAGC RV: ATGGTGCTTGCATTACCACTAA	142
Ls_3765334	$GAG_{(17)}$	FW: GTGGGATGACTAGGTGAGAAGG RV: TATTGTCGATGTGCTCTGCCTA	270
Ls_1127198	$TAT_{(17)}$	FW: CATGTAAACCGAAGTCAACCAT RV: AACTCCTTCATCACCTGAATTG	262
Ls_417667	$AAT_{(16)}$	FW: CAGAGGATCTCATTCCTGTTGA RV: CTCGAAGGCCAATTAGGATAAA	238
Ls_1664757	$TCT_{(16)}$	FW: TATTGTACGGCCCTGTCTGAGT RV: TCCCACCAAGAAGAGAAAGAGA	144
Ls_761180	$CAT_{(16)}$	FW: ATCCGCATTGTTATTGGCAT RV: CTTGTCCTTCTTCCATGCTTCT	242
Ls_1294905	$TTC_{(16)}$	FW: TTCTTCTTCCCTCCATCTTTTG RV: GGAGAGGTTAGGGCACATACAC	230
Ls_1513987	$ATT_{(16)}$	FW: GTTTGACTCGGTTGTGACCTTA RV: GGGTTTCTTTCCTTTTGAACTC	272
Ls_2221540	$TCT_{(16)}$	FW: CACTCATGTGGTTGAGAAGAGC RV: AATCCTTTGGTGAGAAGATTGC	273
Ls_3272325	$GAG_{(16)}$	FW: CTTAGCTCTGTTTCGATGATTGTC RV: GATTTTCCTCACTCTCCAATGC	259
Ls_3901474	$TAG_{(16)}$	FW: GGGCTATGTACCCTTTGACTTTT RV: AAGAATTGTGCAACCACATCAG	231
Ls_254200	$AAT_{(15)}$	FW: TTGTAATGCACCTTCAACTCCA RV: ACCATTTTGGGCAGTGTTTG	252
Ls_752897	$AAG_{(15)}$	FW: GCAGATGTTGATACAGTGGAGG RV: CAGCTTCATTTCGTGGCTAAT	252
Ls_951631	$ATC_{(15)}$	FW: GAAACATCGTGACTTCCTTTGA RV: CAACGAAACAATGGCTACAAAC	150
Ls_954234	$ATT_{(15)}$	FW: TGGCGTTTGGCTAAGTTGTAA RV: GGTTGATTTATGTGTGTGTATGTGG	202
Ls_840190	$TAC_{(15)}$	FW: GATTCTAGCAACTATCACACATGGA RV: ATGTCCCTATCCAAAAGAAAGC	232
Ls_896374	$CAT_{(15)}$	FW: TCCTCTCTACCACTAGACCTAATTGAC RV: TGCATGGTTTATTCTCTCATGG	172
Ls_1743538	$TAA_{(15)}$	FW: GTCTTTTCATCCCAAAGTTCCA RV: TGCCTCAACATCTCCATACAAA	274
Ls_1911570	$GAG_{(15)}$	FW: AATTCCTTCTGGATGGGGTATT RV: ATTAGAGTCATCACCGTCGTCC	278
Ls_3075816	$ATA_{(15)}$	FW: ACACACTTTTGCATGAATGGAG RV: AGAAGGATGTCTGCTCGAAGAT	215
Ls_3452564	$ATT_{(15)}$	FW: CTCGGTTTGTTGTTGTTGTTGT RV: GTTTCATTTGGTCAATCAGTGG	165
Ls_8000298	$ATT_{(15)}$	FW: CCGGTCAACATAAGTCAAATGAG RV: ATATGGAGAAGCCAGCCCC	246

Продолжение Таблицы Д.1

Локус	Мотив	Последовательности праймеров	Размеры продукта
Ls_4591490	$ATA_{(15)}$	FW: GGTCATGCCAATAATGAGAGTG RV: AATTGAACTCGTGGCCTCTAAA	214
Ls_4569173	$TTC_{(15)}$	FW: AGCTCGCACCTCTAAAGAACAA RV: TGGGAATTGTGGTCATATCAAG	240
Ls_1008427	$ATAG_{(13)}$	FW: CACCCCTATCCCACAAATCTTA RV: ATTTATCTTTGGCCCTCATGC	181
Ls_1898261	$ACAT_{(13)}$	FW: CCGAAGGTAGGTTTACAAGCAC RV: CAAACACACACACACACACACA	191
Ls_1524449	$ATAG_{(13)}$	FW: CGACAACACAGTCCATTTCATC RV: ACATCTATTCCCCTCCCAATTC	179
Ls_980491	$CTAT_{(12)}$	FW: TAAGCATTGAGCCTTACAACCA RV: TATAGGGAAAAGGGGAATCTCA	228
Ls_2672894	$TTTG_{(11)}$	FW: CAAAGGATGGAATGTGTCTCAA RV: GTTGGTATGGTTTCCCAGAGTG	163
Ls_65728	$CATA_{(11)}$	FW: ACTAGCTTTCGCGGTTTTCC RV: TATAGTGTCTACGGCTGGGTGG	262
Ls_2552367	$CTAT_{(10)}$	FW: AAAGGTGCAATCACGTAAAGAC RV: ATCGAAGCGGAAAATGTGTA	280
Ls_3952800	$TATG_{(10)}$	FW: AGGATAGAGTGCAGAAGGGTGA RV: AGAGCATTGGATCACTTTGGAT	252
Ls_291006	$AAAG_{(10)}$	FW: TTGAAGGTGGGTTTTACCATTC RV: TAATACTGGATGCACCGTTGAG	168
Ls_648469	$TATG_{(10)}$	FW: GTGATCCATGTTTATGGGGTTA RV: CTCTGGTGTGCCTCTTTTACTG	268
Ls_1512487	$TAGA_{(10)}$	FW: CGATGAGAGGAGAAGAGGTTTT RV: AGTACACTTGACCCATCTTACTGC	277
Ls_1639381	$TATG_{(10)}$	FW: AAGACTCAAGCAATGTGACTAACG RV: ACAGAACCAAATTCTCAATCCC	198
Ls_127910	$TTATC_{(8)}$	FW: CTAGGGAAGAGGAGAAAGGGAT RV: TGAGGGGTAACATAGTCAATCG	262
Ls_897755	$ACCAT_{(7)}$	FW: TGGATGCAAATGAAGGAGTTC RV: AGCAATGGAAAATGGACTAGGA	265
Ls_1453099	$CCCTA_{(7)}$	FW: AGTGGAATAGTTGAGTGGTCACA RV: TCTTGGGAAGGATAGGGTTATG	255
Ls_1588893	$TTTCT_{(7)}$	FW: AAATCCCTCCATCAAGTGTGTC RV: TCCCTATCACCATAGTTAGGCG	275
Ls_556114	$GATAG_{(7)}$	FW: GGTTAGCGAGAAACTAATGGTAGTG RV: ATAGCAATCCCTGTCCTAGCC	261
Ls_4367625	$TTTAT_{(7)}$	FW: TGTTTGAGTTGTCTAGCCTTGG RV: GGTCAATGGGACGAGAAGATAC	222
Ls_4581133	$AGATA_{(7)}$	FW: ACTTCTGGTTTGGCAGAGTTTC RV: TCTTTTGTCATTTGTGGGCA	260
Ls_7621662	$AGATA_{(7)}$	FW: ACCCTAGAACCAAGTCACCTCA RV: AGTTTCTGTTTGATTGCCGACT	248
Ls_77608	$ATGTA_{(7)}$	FW: CCCCTTGTGCATACCATGTAA RV: GACCGACCAATGCTAATCTTTC	263
Ls_143980	$AGATA_{(7)}$	FW: CCTATGGGGTAGTTAATCCACG RV: TCCTACATACTGGAGCAGCAGA	261
Ls_934362	$ATGTA_{(15)}$	FW: CCGAATGAGACACCAGATGATA RV: AGCTCCACTACCTTTGCCTCTA	271

Окончание Таблицы Д.1

Локус	**Мотив**	**Последовательности праймеров**	**Размеры продукта**
Ls_2472377	$TGGATC_{(7)}$	FW: GAAGTAATCGGCAACAAAATGG RV: ACAGATCCAAATCCAGATCCAG	306
Ls_2520134	$AATTTC_{(7)}$	FW: CCAATTCCCCATTTCCAA RV: GAAAATAAAACCGGCAACG	321
Ls_3541267	$TTTTTC_{(7)}$	FW: GTCTTCTTACGCATTCATTCCT RV: CCAACCACACTATTTCATTTCC	249
Ls_2161695	$TAGTT_{(8)}$	FW: GTAAATTCTAAGCAACGTCGGG RV: AAGGTTCCATGTCCTCTCTTCA	239
Ls_4040657	$TCACTT_{(11)}$	FW: TTCACTTTCACAAACCATCACC RV: ACATCTGGCATTTAACCGAAGT	245
Ls_4875128	$TGGCTC_{(7)}$	FW: GGCTCTCTGGGATTATGGTGTA RV: GAGGTCAAGCGAATTACAAAGG	273
Ls_305132	$GTCGGA_{(7)}$	FW: GCAGAGCCGTTATTCGATCTAT RV: CCCTCGTTTCCTCTTCTGACTA	219
Ls_8778872	$TGTTGA_{(7)}$	FW: TAGGATTCTTCGGTGTGACCTT RV: GCTATCTGGTATGTCCTCTGGG	221
Ls_2104879	$ATATAG_{(10)}$	FW: CCCACTTCACATAAAGATTTGTCC RV: AGCCCCTTGCGGGTATTTA	375

Приложение Д.2

14 локусов, отобранных по результатам тестирования на трех видах лиственницы

Локус	Мотив	Нуклеотидная последовательность праймеров	Размеры продукта	Количество выявленных аллелей
Ls_3765334	$GAG_{(17)}$	FW: GTGGGATGACTAGGTGAGAAGG RV: TATTGTCGATGTGCTCTGCCTA	174-213	5
Ls_1247092(2)	$CTT_{(19)}$	FW: AAACGCCAACCACAGAATTTAC RV: TTATCAAACGAGGGATGCTTCT	201-228	7
Ls_417667	$AAT_{(16)}$	FW: CAGAGGATCTCATTCCTGTTGA RV: CTCGAAGGCCAATTAGGATAAA	207-243	5
Ls_611965	$CAG_{(18)}$	FW: GCAAGTTCCGATTCAAGAGTTT RV: GTTGCTGCTGAGGCATGTAG	222-276	7
Ls_752897	$AAG_{(15)}$	FW: GCAGATGTTGATACAGTGGAGG RV: CAGCTTCATTTCGTGGCTAAT	216-264	12
Ls_840190	$TAC_{(15)}$	FW: GATTCTAGCAACTATCACACATGGA RV: ATGTCCCTATCCAAAAGAAAGC	216-249	7
Ls_954234	$ATT_{(15)}$	FW: TGGCGTTTGGCTAAGTTGTAA RV: GGTTGATTTATGTGTGTGTATGTGG	171-204	10
Ls_3952800	$TATG_{(10)}$	FW: AGGATAGAGTGCAGAAGGGTGA RV: AGAGCATTGGATCACTTTGGAT	200-264	9
Ls_980491	$CTAT_{(12)}$	FW: TAAGCATTGAGCCTTACAACCA RV: TATAGGGAAAAGGGGAATCTCA	204-240	3
Ls_2672894	$TTTG_{(11)}$	FW: CAAAGGATGGAATGTGTCTCAA RV: GTTGGTATGGTTTCCCAGAGTG	152-164	3
Ls_2552367	$CTAT_{(10)}$	FW: AAAGGTGCAATCACGTAAAGAC RV: ATCGAAGCGGAAAATGTGTA	184-196	4
Ls_1008427	$ATAG_{(13)}$	FW: CACCCCTATCCCACAAATCTTA RV: ATTTATCTTTGGCCCTCATGC	152 - 174	5
Ls_4040657	$TCACTT_{(11)}$	FW: TTCACTTTCACAAACCATCACC RV: ACATCTGGCATTTAACCGAAGT	194-218	3
Ls_305132	$GTCGGA_{(7)}$	FW: GCAGAGCCGTTATTCGATCTAT RV: CCCTCGTTTCCTCTTCTGACTA	210-240	6

Приложение Д.3

Число аллелей для 14 микросателлитных локусов, протестированных на 24 образцах четырех популяций лиственниц сибирской, Гмелина и Каяндера.

Локусы	Аллели	ЛС-СЕ	ЛС-ТУ	ЛГ-Ч	ЛК-Я
		24	24	24	24
Ls_980491	204	0,479	0,479	-	-
	236	0,521	0,521	1,000	0,917
	240	-	-	-	0,083
Ls_2672894	152	0,938	0,979	0,875	0,958
	156	-	-	0,083	0,021
	164	0,062	0,021	0,042	0,021
Ls_4040657	194	1,000	1,000	0,958	0,895
	200	-	-	-	0,063
	218	-	-	0,042	0,042
	null			0.167	
Ls_1008427	152	0,458	0,625	0,916	0,896
	160	0,292	0,271	0,042	0,083
	164	-	0,104	-	-
	168	0,146	-	0,042	0,021
	174	0,104	-	-	-
	null			0.229	
Ls_417667	207	-	-	0,021	0,146
	213	-	-	0,021	-
	219	0,021	0,021	0,063	0,021
	228	0,625	0,521	0,895	0,833
	243	0,354	0,458	-	-
Ls_840190	216	-	0,062	-	-
	228	-	-	-	0,146
	237	0,542	0,188	0,250	0,687
	240	0,229	0,729	0,604	0,167
	243	0,208	-	0,146	-
	246	0,021	-	-	-
	249	-	0,021	-	-
	null	0.229			
Ls_954234	171	-	-	0,021	-
	174	-	-	0,021	0,042
	177	0,645	0,521	0,438	0,396
	180	-	-	0,021	-
	183	0,125	0,250	-	0,042
	186	0,063	-	-	0,021
	192	-	-	0,332	-
	195	0,146	0,208	0,104	0,313
	198	-	-	0,021	0,021
	204	0,021	0,021	0,042	0,165
	null	0.104		0.146	

Продолжение Таблицы Д.3

Локусы	**Аллели**	**ЛС-СЕ**	**ЛС-ТУ**	**ЛГ-Ч**	**ЛК-Я**
Ls_752897	216	-	-	0,124	0,122
	225	-	-	0,021	0,042
	228	-	-	0,125	0,188
	231	0,124	-	0,104	0,063
	237	0,042	0,166	0,042	0,063
	240	0,021	0,021	0,167	-
	243	0,188	0,188	0,208	0,292
	246	0,021	-	0,188	0,021
	249	0,604	0,625	-	0,125
	251	-	-	0,021	-
	255	-	-	-	0,042
	264	-	-	-	0,042
	null	0.104		0.188	
Ls_2552367	184	0,854	0,688	0,895	0,750
	188	0,021	0,083	0,021	-
	192	0,104	0,208	0,063	0,208
	196	0,021	0,021	0,021	0,042
Ls_1247092(2)	201	-	-	0,062	-
	207	0,125	0,417	0,229	0,104
	210	0,083	-	0,396	0,021
	216	0,542	0,208	0,188	0,188
	222	-	-	0,021	0,354
	225	0,167	0,375	0,104	0,229
	228	0,083	-	-	0,104
	null	0.313		0.188	0.125
Ls_3765334	174	-	-	0,354	0,625
	177	0,750	0,75	0,625	0,375
	192	0,021	-	-	-
	198	-	0,021	-	-
	213	0,229	0,229	0,021	-
Ls_611965	222	0,708	0,500	0,062	0,125
	231	-	-	0,146	0,083
	237	0,125	0,187	0,729	0,75
	240	-	-	-	0,021
	243	-	-	0,021	-
	261	0,125	0,313	0,042	0,021
	276	0,042	-	-	-

Окончание Таблицы Д.3

Локусы	**Аллели**	**ЛС-СЕ**	**ЛС-ТУ**	**ЛГ-Ч**	**ЛК-Я**
Ls_3952800	200	0,021	-	0,250	0,521
	212	-	-	-	0,061
	216	-	0,188	0,082	-
	224	-	-	0,042	-
	240	-	-	0,354	0,313
	244	0,521	0,333	0,146	-
	252	0,333	0,479	0,063	0,021
	256	0,083	-	0,021	0,063
	264	0,042	-	0,042	0,021
	null	0.375	0.313	0.375	
Ls_305132	210	0,083	0,083	-	0,021
	216	0,104	0,250	-	-
	222	0,521	0,542	0,854	0,874
	228	0,292	0,104	0,125	0,042
	234	-	0,021	-	0,042
	240	-	-	0,021	0,021
	null		0.250		