На правах рукописи

Бондар Евгения Ивановна

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

1.5.7 – Генетика

АВТОРЕФЕРАТ
Диссертация на соискание учёной степени
кандидата биологических наук

Красноярск 2023

Работа выполнена на кафедре геномики и биоинформатики Федерального государственного автономного образовательного учреждения высшего образования «Сибирский федеральный университет», г. Красноярск

Научный руководитель:	**ТАТАРИНОВА Татьяна Валерьевна** PhD, доцент кафедры биологии Университета Ла Верна, заведующий кафедрой вычислительной биологии Флетчера Джонса, Ла Верн, США
Официальные оппоненты:	**СЕМЕРИКОВ Владимир Леонидович** доктор биологических наук, заведующий лабораторией молекулярной экологии растений Федерального государственного бюджетного учреждения науки Институт экологии растений и животных Уральского отделения Российской академии наук, г. Екатеринбург **РАЙКО Михаил Петрович** кандидат биологических наук, старший научный сотрудник лаборатории «Центр биоинформатики и алгоритмической биотехнологии» Федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет», г. Санкт-Петербург
Ведущая организация:	Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», г. Новосибирск

Защита состоится «__» _____ 20__ г. в ____ часов на заседании диссертационного совета Д.002.214.01 в Федеральном государственном бюджетном учреждении науки Институт общей генетики им. Н.И. Вавилова Российской академии наук по адресу: 119991, Москва, ул. Губкина, 3.
Тел: (499) 135-62-13, факс: (499) 132-89-62, e-mail: dissovet@vigg.ru. С диссертацией и авторефератом можно ознакомиться в библиотеке и на сайте Института www.vigg.ru.

Автореферат разослан «___» _____2024 года.

Ученый секретарь диссертационного совета,
доктор биологических наук Горячева И. И.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

У более чем 70% всех видов растений не было секвенировано ни одного участка ДНК, не говоря уже о полном геноме (Kress et al. 2022). Из геномов 1310 уникальных 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 (“NCBI Genome [Internet]” 2023). Таксономическое распределение публично доступных геномов растений довольно смещено в сторону сельскохозяйственных культур (Lughadha et al. 2016). Многие опубликованные геномы, хотя и достаточно полные на уровне последовательности, имеют очень фрагментарные сборки. Обилие псевдогенов, увеличенное число генных семейств (Qiao et al. 2019) и пролиферирующая активность мобильных элементов (Pellicer et al. 2018; Claros et al. 2012) 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 (McLoughlin 2021; Brenner and Stevenson 2006). 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 (Sun et al. 2022; Niu et al. 2022; Mosca et al. 2019; Kuzmin et al. 2019; A. V. Zimin et al. 2017; Neale et al. 2017; Gonzalez-Ibeas et al. 2016; Warren et al. 2015; Nystedt et al. 2013). 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 (Абаимов 2009). Этот вид отличается высокой устойчивостью к низким температурам и гниению древесины, а также быстрым ростом, что делает его особенно ценным для использования в строительстве. Хотя геном лиственницы сибирской был впервые опубликован в 2019 году (Kuzmin et al. 2019), генная аннотация была выполнена только в представленной к защите работе и сделана публично доступной (Bondar et al. 2022a,b).

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.) (Nystedt et al. 2013), ель белую (*P. glauca)* (Warren et al. 2015), сосну ладанную (*Pinus taeda* L.) (Neale et al. 2014; A. Zimin et al. 2014; A. V. Zimin et al. 2017), сосну сахарную (*Pinus lambertiana* Douglas) (Stevens et al. 2016), псевдотсугу Мензиса (*Pseudotsuga menziesii* (Mirb.) Franco) (Neale et al. 2017), пихту белую (*Abies alba* Mill.) (Mosca et al. 2019), лиственницу сибирскую (*Larix sibirica),* лиственницу японскую (*Larix kaempferi* (Lamb.) Carr.) (Sun et al. 2022), сосну красную китайскую (*Pinus tabuliformis* Carr.) (Niu et al. 2022), 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 её популяционной структуры (Dulamsuren et al. 2010; Semerikov et al. 2013; Tumenjargal et al. 2020) и разработку молекулярно-генетических маркеров (Babushkina et al. 2016; Oreshkova, Belokon, and Jamiyansuren 2013). Полногеномное

секвенирование предоставило геномные ресурсы и позволило разработать дополнительные высокоинформативные видоспецифичные SSR (simple sequence repeat)-маркеры *L. sibirica* (Oreshkova et al. 2017; 2019), которые можно использовать в различных практических приложениях, в том числе для отслеживания происхождения древесины при борьбе с незаконными рубками (Krutovsky et al. 2019). Публикации первых ядерного (Kuzmin et al. 2019), хлоропластного (Bondar et al. 2019) и митохондриального (Putintseva et al. 2020) геномов лиственницы сибирской, а недавно и лиственницы японской (Sun et al. 2022), значительно способствовали развитию геномного ресурса для рода *Larix* и хвойных в целом.

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 популяционных исследований лиственниц Гмелина и Каяндера. Для трёх 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.

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 также лиственниц Гмелина и Каяндера и специально разработаны с учётом их возможного использования

также 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 работе, включая файлы аннотации и комплексную библиотеку повторов, доступны публично на платформе 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. Набор из 14 полиморфных микросателлитных маркеров, разработанных в данном исследовании для лиственницы сибирской, может также использоваться для популяционно-генетических исследований лиственницы Гмелина и Каяндера.

Апробация результатов исследования

Доклады по теме диссертации проводились на ежегодных заседаниях кафедры геномики и биоинформатики СФУ в 2018-2022 гг. 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 г., 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 работы

Материалы диссертации представлены в 5 статьях, опубликованных в международных рецензируемых изданиях, индексируемых в базах Web of Science и Scopus, а также в 14 тезисах международных и всероссийских конференций.

Личный вклад автора в проведенные исследования

Автором 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 под руководством Орешковой Н.В. в рамках гранта Правительства РФ (договор № 14.Y26.31.0004, руководитель проекта проф. К.В. Крутовский). 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 (руководитель проекта проф. Крутовский К.В.), 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 программного конвейера 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. Тестирование микросателлитных маркеров проводилось под руководством Орешковой Н.В. 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 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 перед ее апробацией.

Автор благодарен сотрудникам отдела мониторинга состояния лесных генетических ресурсов Центра защиты леса г. Красноярска за предоставленные образцы лиственницы сибирской.

Также автор признателен Мирошниковой К.А., Бирюкову В.В., Акуловой В.С., Новиковой С.В. и Тараненко Е.А. за 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.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Глава 1. Обзор литературы

В разделе 1.1 дан краткий обзор доступных на сегодняшний день геномов и геномных аннотаций зеленых растений и хвойных видов, описаны особенности геномов хвойных и сложности работы с ними. В разделе 1.2 дана краткая характеристика вида лиственницы сибирской, его экологического ареала и хозяйственной ценности. В разделе 1.3 описан общий подход к геномной аннотации, перечислены основные этапы и наиболее распространенные ошибки, возникающие в ходе получения аннотации, описаны методы поиска и маскировки повторов, перечислены особенности *ab initio* предсказания генных моделей, обсуждаются преимущества и недостатки использования данных РНК-секвенирования для повышения точности предсказания генов, перечислены основные ресурсы и методы для функциональной аннотации генов. В разделе 1.4 описана типичная структура промоторной области в растительном геноме, перечислены наиболее распространенные методы полногеномного обнаружения сайтов начала транскрипции, описаны некоторые статистические свойства промоторных областей и кодирующий частей генома, таких как GC3 и GC-skew. В разделе 1.5 описана методика разработки микросателлитных маркеров для изучения генетического разнообразия растений и проведения SSR анализа.

Глава 2. Материалы и методы

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 (Kuzmin et al. 2019). Для обеспечения поддержки предсказанных генных моделей были использованы референсные транскриптомы на основе пяти тканей — почки, хвои, камбия, проростка и побега первого года, взятых от референсного дерева лиственницы сибирской (NCBI: GIXH00000000, GJYD00000000, GJYL00000000, GJYN00000000 и GJYW00000000).

Анализ и маскирование высокоповторяющихся элементов. Для поиска мобильных элементов *de novo* был использован RepeatModeler (Smit and Hubley 2008). Для маскирования областей низкой сложности и повторов был использован RepeatMasker (Smit, Hubley, and Green 2013) в сочетании с комбинированной библиотекой повторов. Библиотека RepeatModeler была дополнена кластеризацией часто встречающихся прочтений из данных полногеномного секвенирования, нераспознанные элементы которой были классифицированы с помощью TEclass (Abrusán et al. 2009), а также библиотеками повторов RepBase 2017.01.27 (Bao, Kojima, and Kohany 2015), MIPS (Nussbaumer et al. 2013), CPRD (Wegrzyn et al. 2013) и PIER (Neale et al. 2014; Wegrzyn et al. 2013). Часть длинных прочтений Oxford Nanopore, доступная для генома лиственницы сибирской, также использовалась для оценки численности повторов. Для поиска тандемных повторов были использованы программы GMATo (X. Wang, Lu, and Luo 2013) и TRF (Benson 1999).

Оценка времени вставки ретротранспозонов LTR-RT. Для *de novo* идентификации LTR-RT элементов был использован LTRharvest (Ellinghaus, Kurtz, and Willhoeft 2008), потенциальные ложноположительные совпадения были отфильтрованы с помощью LTR_retriever (Ou and Jiang 2018). Полученные результаты были объединены с библиотекой LTR-RT элементов созданной Zhou с соавторами (Zhou et al. 2021). Время эволюционного расхождения LTR-RT последовательностей (T) после дуплицирования (вставки) было рассчитано с использованием генетического расстояния (d) по модели Jukes-Cantor (Jukes and Cantor 1969):

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

где μ — частота мутаций и p — доля различий между последовательностями ($p = 1$ – идентичность, где значения идентичности аппроксимируется по результатам сравнения нуклеотидных последовательностей LTR-RT внутри вида с помощью blastn). Время расхождения выражается в миллионах лет, при скорости синонимичных замен $\mu = 1{,}57 \times 10^{-8}$ на сайт в год (De La Torre et al. 2017).

Аннотация с использованием программы MAKER2. MAKER2 (Holt and Yandell 2011) использовался для получения структурной аннотации полного генома. Для поиска предварительных генных моделей использовался AUGUSTUS (Stanke et al. 2006). Данные транскриптома лиственницы сибирской и общедоступные сборки транскриптов родственных видов хвойных (*Gnetum gnemon, Picea abies, Pinus lambertiana, Podocarpus macrophyllus, Pseudotsuga menziesii* и *Pinus taeda*), использовались в качестве вспомогательных данных. Uniprot использовался в качестве эталонной базы данных белков.

Оценка полноты сборки и функциональная аннотация. Для картирования белков, полученных в ходе структурной аннотации, использовалась база NCBI GenBank nr, отфильтрованная по идентификатору таксономии на уровне Embryophyta. Также все предсказанные гены были картированы на полную базу NCBI GenBank nr с использованием blastp; совпадения с бактериями, грибами и археями были удалены (e-value < $1{\times}10^{-5}$, процент совпадений > 20, доля покрытых high-scoring pair > 20). Поиск белковых доменов проводился с помощью InterProScan (Blum et al. 2021). Картирование терминов GO было выполнено с использованием Blast2GO OmixBox (Conesa and Götz 2008; Götz et al. 2008). Для выявления терминов GO, в которых число картированных генов значительно различается, использовался тест пропорций. Для коррекции *p*-значений с целью контроля ложноположительных результатов были использованы два метода, расчёт FDR (Benjamini and Hochberg 1995) и (Storey 2002).

Аннотация органельных геномов лиственницы сибирской

Сборка и аннотация хлоропластного генома. В качестве референса для сборки и аннотации были использованы хлоропластные геномы *Larix decidua* Mill. и *L. occidentalis* Nutt. (NCBI Genbank AB501189.1 и FJ899578.1, соответственно). Прочтения полногеномного секвенирования были картированы на референсные геномы хлоропластов *L. decidua* и *L. occidentalis* с помощью Bowtie2 (Langmead and Salzberg 2012). Выровненные прочтения были собраны с помощью ассемблера SPAdes. Скаффолдинг выполнялся с использованием MP-прочтений (Mate Pair, парные прочтения с длинной вставкой) с помощью SSPACE (Boetzer et al. 2011). Для аннотации использовался сервис Rapid Annotation with Subsystem Technology (RAST) (Overbeek et al. 2014).

Сборка и аннотация митохондриального генома. Выделение интактных митохондрий и получение обогащенной мтДНК проводилось на базе лаборатории генетической инженерии растений Сибирского института физиологии и биохимии растений (СИФИБР СО РАН) под руководством Ю.М. Константинова. Секвенирование и первичная обработка прочтений проводились сотрудниками лаборатории лесной

геномики СФУ под руководством проф. К.В. Крутовского. Гибридная сборка с использованием длинных прочтений MinION (Oxford Nanopore Technologies, Inc., Оксфорд, Великобритания) и коротких прочтений PE Illumina проводилась с помощью MaSuRCA (Zimin et al. 2013) под руководством Д.А. Кузьмина, В.В. Шарова и Ю.А. Путинцевой (A. V. Zimin et al. 2013). Оценка точности гибридной сборки проведена с помощью REAPR (Hunt et al. 2013). Для поиска повторов в сборке митогенома был использован RepeatModeler. Неизвестные повторы в *de novo* библиотеке классифицированы с помощью TEclass. Так же были использованы дополнительные библиотеки RepBase (Bao, Kojima and Kohany 2015), MIPS (Nussbaumer et al. 2013), CPRD (Wegrzyn et al. 2013) и PIER (Neale et al. 2014). Гены тРНК проаннотированы с помощью ARAGORN (Laslett and Canback 2004) и tRNAscan-SE (Lowe and Eddy 1997). Рибосомальные РНК (рРНК) проаннотированы с помощью RNAmmer (Lagesen et al. 2007). Белок-кодирующие гены проаннотированы с помощью BLAST.

Предсказание TSS. В дополнение к *L. sibirica* были использованы аннотации геномов *P. taeda*, *P. glauca* и *P. abies.* Предсказание TSS было выполнено на последовательностях, определенных как области –1000 и +250 п.н. вокруг стартового кодона каждого гена, с использованием программы TSSPlant (Shahmuradov, Umarov, and Solovyev 2017). Для выбора наилучшего предсказания для каждого гена использовалось распределение длин 5'-нетранслируемых областей (5'-UTR). Основываясь на предположении, что длина 5'-UTR может быть описана гамма-распределением, был составлен пул длин 5'-UTR на основе аннотаций нескольких модельных растений (*Arabidopsis thaliana* (L.) Heynh., *Oryza sativa* L., *Sorghum bicolor* (L.) Moench и *Populus trichocarpa* Torr. & A.Gray ex. Hook.). Функция плотности вероятности была применена для выбора наиболее вероятных положений TSS.

Анализ нуклеотидного состава промоторов и генов. Анализ частоты нуклеотидов в промоторах был проведен на последовательностях, центрированных по позиции предсказанных TSS (–1000, +200 вокруг TSS). Частота мотивов CA и TATA была рассчитана в скользящем окне с использованием пакета stringr R. CG-skew последовательности определялся как пропорция (C-G)/(C+G) и рассчитывался в скользящем окне вдоль последовательности промотора. GC_3 рассчитывался на кодирующих последовательностях генов с помощью R пакета seqinr. Изменение стандартной свободной энергии ДНК оценивалось с помощью PromPredict (Rangannan and Bansal 2010) в скользящем окне. Поиск сайтов связывания транскрипционных факторов выполнялся с использованием базы данных TRANSFAC.

Разработка и апробация микросателлитных маркеров. Отбор контигов, содержащих микросателлитные повторы, проводился при помощи GMATo. Минимальное число повторений мотива для трехнуклеотидных — 15 раз, для четырехнуклеотидных — 10, для пяти- и шестинуклеотидных — 7 раз. Дизайн праймеров для отобранных повторов проводился в онлайн-сервисе WebSat. Праймеры были отфильтрованы, чтобы исключить неуникальные и принадлежащие хлоропластному и митохондриальному геному. Отбор полиморфных маркеров и их тестирование проводились на выборках из 4 популяций *L. sibirica* (Северо-Енисейский район Красноярского края и окрестности села Туим, Республика Хакасия), *L. gmelinii* (окрестности населенного пункта Хилок, Забайкальский край) и *L. cajanderi* (Намский район, Республика Саха (Якутия)). Для проведения ПЦР использовали готовые реакционные смеси для амплификации ДНК «GenePak PCR Core» производства ООО «Лаборатория Изоген» (Москва, Россия). 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. Показатели, отражающие гетерозиготность, степень генетической подразделенности, межпопуляционной дифференциации и уровень

дивергенции были рассчитаны при помощи программного обеспечения GenAlEx 6.5 (Peakall and Smouse 2012).

Глава 3. Результаты и обсуждение

Анализ высокоповторяющихся элементов генома. Отдельная видоспецифичная *de novo* библиотека повторов RepeatModleler, а также комбинированная библиотека, использованная для идентификации повторов в геноме лиственницы, депонированы в сервисе figshare с DOI 10.6084/m9.figshare.19785913, а также могут быть найдены по адресу https://hpccloud.sfu-kras.ru/owncloud/index.php/s/Dlb3rGQD2esWQvW.

Относительное содержание классифицированных семейств повторов, обнаруженных в геноме лиственницы сибирской, аналогично описанному ранее 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. Доля генома, покрытая повторами в части длинных ридов Oxford Nanopore, по оценке RepeatMasker, составила 66%. Класс I длинных концевых ретротранспозонов (LTR), представленных в основном элементами Gypsy и Copia, составляет наибольшую часть всех мобильных элементов, что широко наблюдается также у других хвойных (Nystedt et al. 2013; Wegrzyn et al. 2014; Neale et al. 2017; Perera et al. 2018) и покрытосеменных (Civáň et al. 2011). Значительная часть LTR была гомологична LTR в фрагментах сосны ладанной в отсеквенированных библиотеках искусственных бактериальных хромосом (Magbanua et al., 2011; Wegrzyn et al., 2013). Среди non-LTR-ретротранспозонов LINE/L1, I, Penelope и SINE вместе составляют около 98% всех non-LTR-ретротранспозонов, которые покрывают 12% длины сборки (**Рисунок 1А**). DNA транспозоны класса II покрывают около 5% размера сборки, из них 4,5% не были классифицированы с помощью TEclass. Среди классифицированных транспозонов наиболее многочисленны терминальные инвертированные повторы TIR (0,16% DNA транспозонов), Helitron (0,06%), EnSpm (0,02%), hAT (<0,01%) (**Рисунок 1А**).

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 268,7 локусов на миллион п.н. По сравнению с другими видами сборка генома лиственницы также имела относительно высокую плотность SSR, аналогичную геномам ели европейской и тополя черного (**Рисунок 1В**). В работах (Wegrzyn et al. 2014) и (Neale et al. 2014) сообщалось о плотности SSR 10-20 локусов/Mbp для *Pinus taeda, Picea abies* и *Picea glauca*, при оценке с помощью TRF. Мы также просканировали геном лиственницы с помощью TRF, что дало 17 145 локусов с тем же размером мотива и с общей плотностью 4,1 локуса/Mbp.

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. Однако доля генома, представленного простыми повторами и мобильными элементами в текущей сборке генома составила только 40%.. При этом доля покрытой повторами порции длинных прочтений Oxford Nanopore, по оценкам RepeatMasker, составила 66% п.н., что предполагает, что обогащенная повторами часть генома лиственницы сибирской была слишком фрагментирована, чтобы быть включенной в окончательную сборку. Эта оценка тем не менее ниже, чем у всех остальных голосеменных растений. Похожие данные были в 2021 году получены для двух других видов лиственницы, *Larix decidua* и *Larix kaempferi* (Heitkam et al. 2021), что может говорить о том, что сравнительно меньшая доля повторов в геноме характерна для видов рода *Larix,* и в целом согласуется с ролью транспозонов и мобильных элементов ДНК в увеличении размера генома хвойных.

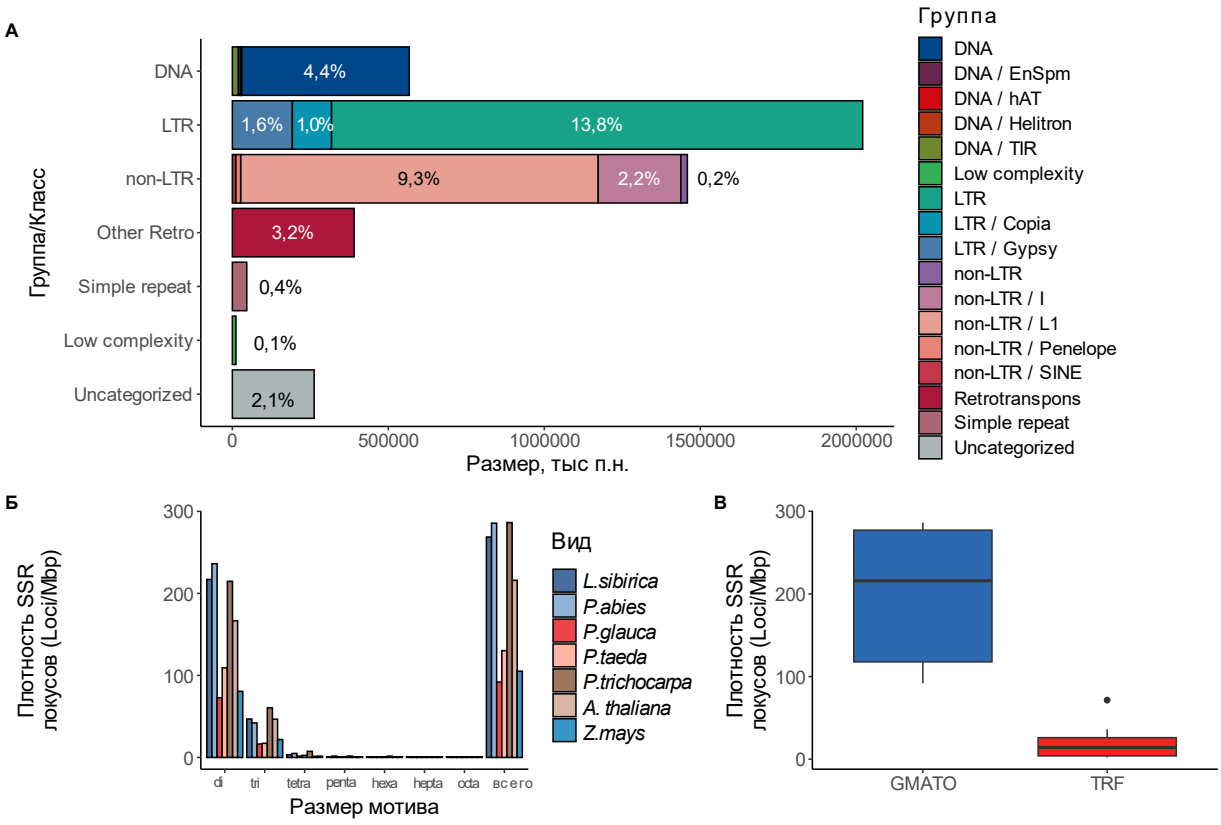

Рисунок 1. А – относительное содержание семейств повторов в геноме лиственницы сибирской; **Б –** плотность микросателлитных локусов (число локусов с ди-, три-, тетра-, пета-, гекса-, гепта- и октонуклеотидными мотивами на 1 млн. п.н.) для нескольких видов хвойных и покрытосеменных видов, оцененная с помощью GMATo (Larch – *Larix sibirica*, Pab – *Picea abies*, PG – *Picea glauca*, Pita – *Pinus taeda*, Popul – *Populus trichocarpa*, TAIR - *Arabidopsis thaliana*, Zea – *Zea mays*); **В –** среднее число микросателлитных локусов найденное у всех видов, перечисленных в **Б**, с использованием GMATo и TRF.

В литературе обсуждается причина и механизм накопления повторов в геномах покрытосеменных и хвойных (Nystedt et al. 2013; Pellicer et al. 2018). Мнения по этому поводу противоречивы. Некоторые исследователи считают это результатом всплесков активности мобильных элементов (Belyayev 2014; Naville et al. 2019; Piegu et al. 2006; Tsukahara et al. 2009; Zeh, Zeh, and Ishida 2009), в то время как другие предполагают, что большие геномы, содержащие множество разнообразных повторов, могли приобретать их с течением времени в результате постоянного процесса накопления, а не подвергаться внезапной экспансии определенного повторяющегося элемента. Это может также означать, что процесс элиминации повторов может быть более медленным или менее эффективным, что приводит к замедлению сокращения размера генома (Kelly et al. 2015; Pellicer et al. 2018; W. Wang et al. 2016).

Оценка времени вставки ретротранспозонов LTR-RT. Ретротранспозоны класса I размножаются за счет интеграции своей промежуточной РНК в геном хозяина посредством ретротранскрипции в кДНК, используя механизм транскрипции хозяина и собственные ферменты. Когда вставка ретроэлемента только что произошла, два фланкирующих LTR на его 5'- и 3'-концах идентичны (Kumar and Bennetzen 1999), но со временем они накапливают мутации, и, что особенно важно, частота этих мутаций, как правило выше, чем в кодирующих областях, поскольку повторы в отличие от генов не находятся под давлением отбора. Количество различий в последовательностях между двумя фланкирующими LTR можно использовать в качестве приближения для оценки времени, когда элемент встроился в геном. Оценка времени встраивания элементов LTR

может пролить свет на эволюционные аспекты организации генома и потенциально обнаружить недавние и древние события повторной экспансии ретротранспозонов.

LTRharvest с последующей фильтрацией LTR_retriever идентифицировали 347 LTR элементов и 36 интактных LTR в сборке лиственницы сибирской. Эти 36 интактных LTR были объединены с 367 элементами, идентифицированными Zhou et al. 2021. На основе оценки количества замен в фланкирующих частях 403 LTR элементов, вероятная волна встраивания ретротранспозонов в геном лиственницы скорее всего произошла порядка 4-5 млн лет назад (**Рисунок 2В**). Хотя суперсемейства Copia (PR-INT-RT) и Gypsy (PR-RT-INT) (**Рисунок 2А**) имеют немного разные профили, их средние и медианные значения очень 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 2Г**). По сравнению с профилями других голосеменных, лиственница показала самый древний всплеск LTR, даже по сравнению с гнетумом и гинкго (**Рисунок 2Б**).

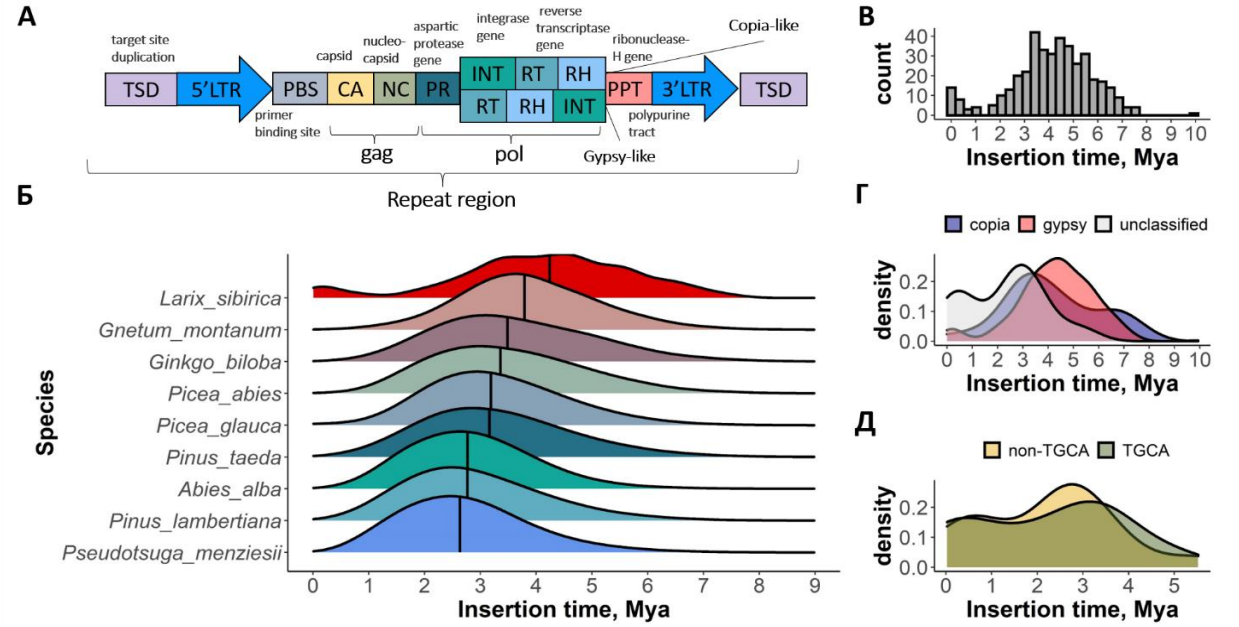

Рисунок 2. А – структура Copy- и Gypsy-подобных LTR ретротранспозонов; **Б –** оценка времени вставки LTR элементов в геноме девяти покрытосеменных; **В –** оценка времени вставки LTR элементов в геноме лиственницы сибирской; **Г** – оценка времени вставки суперсемейств Copia и Gypsy; **Д –** оценка времени вставки TGCA/non-TGCA LTR элементов. По оси абсцисс – время вставки в млн. лет.

Типичные оценки времени вставки LTR в геномы растений варьируются от 1 до 2,5 млн лет назад для покрытосеменных (Brunner et al. 2005; Paterson et al. 2009; Buti et al. 2011; Zhao et al. 2013; Yin et al. 2015). Сообщалось, что у голосеменных растений время встраивания оценивается в 10–15 млн лет назад (Wan et al. 2021). На основании идентификации LTR, проведенной Zhou с соавторами (Zhou et al. 2021), приблизительное время экспансии LTR голосеменных можно оценить в 2–4 млн лет назад. На меньшее число повторов у лиственницы, по сравнению с другими хвойными, 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 оценка времени встраивания LTR для них также сходна, несмотря на заметное различие в числе идентифицированных LTR (403 у лиственницы, 31 016 у ели европейской и 34 952 у пихты). Таким образом, меньшее число повторов у лиственницы по сравнению с другими

хвойными вероятно объективно и не может быть объяснено только фрагментарностью сборки.

Структурная аннотация с использованием программы MAKER2. Структурная аннотация полногеномной сборки с помощью пайплайна MAKER2 на кластере из 448 ядер заняла 22 дня, исключая настройку AUGUSTUS и подбор базы данных повторов. Было получено 39 370 генных моделей, состоящих из 134 271 экзона и 94 901 интрона (**Таблица 1**). В качестве показателя контроля качества MAKER2 использует расстояние между разными версиями («редакциями») аннотаций (annotation edit distance, AED) (Holt and Yandell 2011). Для аннотации лиственницы сибирской AED, рассчитанное MAKER2, было ниже 0,5 для 95% моделей генов, что сравнимо с таковым для генома мыши GRCm37 (Holt and Yandell 2011). Среди 39 370 генов 77% были подтверждены либо данными секвенирования РНК из нескольких тканей, либо гомологией с генами в базе NCBI nr (taxID *Embryophyta*).

Таблица 1. Статистика геномной сборки и аннотации

Параметр	***Larix sibirica***
Количество хромосом	12
Размер генома, Gbp	12,03[а]
Размер сборки, Gbp	5,59 [б] / 12,34 [в]
N50, bp	3098 [б]/ 6443 [в]
GC состав, %	35,41
Содержание повторов, %	66
Количество предсказанных генных моделей	39370
Количество полно-длинных генных моделей	24551
Средняя длина CDS, bp	244,29
Средняя длина интрона, bp	360,93
Максимальны длина интрона, bp	10153

[а] Pellicer and Leitch, 2020

[б] Контиги

[в] Скаффолды

Для областей, идентифицированных RepeatMasker как повторы, также были обнаружены пересечения с кодирующими частями предсказанных моделей генов. Всего 6 884 гена имели по крайней мере 20% перекрытия с повтором (**Рисунок 3А**). 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, перекрывающихся с повторами, были рецептороподобные протеинкиназы, богатые лейциновыми повторами белки (leucine-rich repeats), факторы транскрипции, АТФ-связывающие переносчики, ферменты синтазы, редуктазы, эстеразы и пероксидазы, ферменты цитохрома С и цитохрома Р450.

Так же как и размеры генома, средняя длина интрона больше у хвойных, чем у покрытосеменных растений (Sena et al. 2014). Максимальная длина интрона в геноме лиственницы составила 10153 п.н., что меньше, чем у других видов хвойных. 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 3**). 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. (12,9% интронного пространства) покрыты повторами.

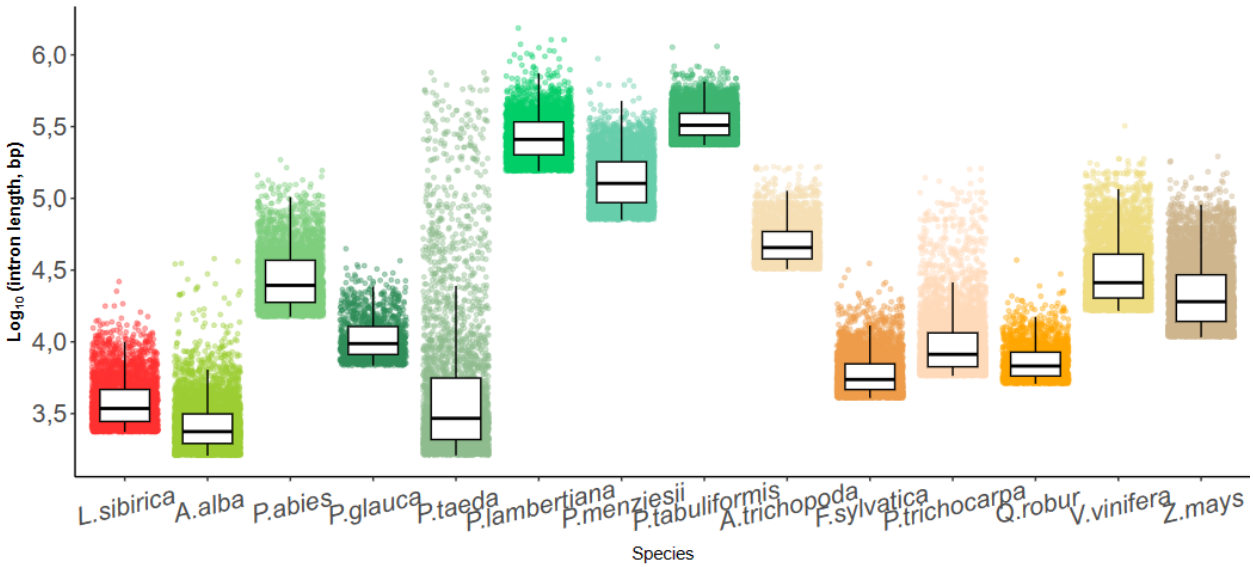

Рисунок 3. 10% самых длинных интронов в 11 видах покрытосеменных и хвойных растениях.

Функциональная аннотация. 87% предсказанных генных моделей лиственницы (34 358 из 39 370) имели гомологию с белками *A. thaliana*, (минимальное e-value 10^{-5}, минимальное покрытие 20% и минимальная идентичность 20%) (**Рисунок 4**). 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 4Б**). При обратном картировании 72% белков *A. thaliana* (19 706 из 27 416) имели гомологию среди белков лиственницы.

Терминами GO было проаннотировано 30 512 генных моделей (78%). Все гены были разделены на 20 функциональных категорий, функции классифицированы в соответствии с последней версией словаря генной онтологии: 5 категорий в «биологические процессы», 6 в «молекулярные функции», 5 в «клеточные компоненты». Все белки из соответствующей категории были картированы на базу белков *A. thaliana* 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 *A. thaliana*.

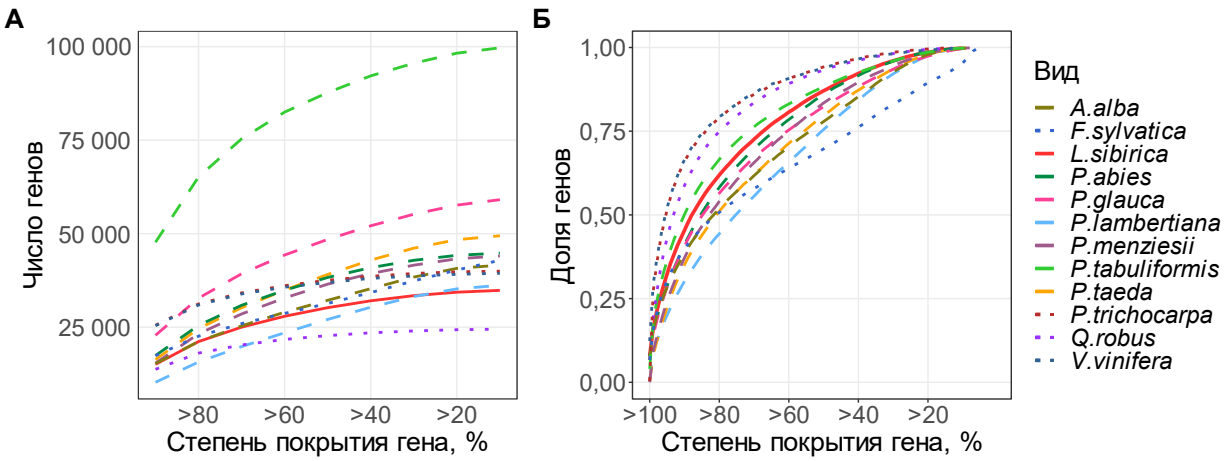

Рисунок 4. Кумулятивное количество (**А**) и доля (**В**) генов, имеющих сходство с белками *Arabidopsis* (покрытие *qcovhsp* выше указанного порога, минимальная идентичность 20%). Голосеменные отмечены штрих-линией, покрытосеменные — пунктиром, лиственница сибирская – сплошной линией.

Органельные геномы лиственницы сибирской

Хлоропластный геном. Общая длина финальной сборки хлоропластного генома составила 122 560 п.н., полученная сборка была депонирована в GenBank NCBI (NC_036811.1). Аннотация с помощью RAST и сравнение с имеющимися аннотациями для *L. decidua* и *L. occidentalis* позволили идентифицировать 110 генов, из которых 34 представляют собой гены тРНК, 4 — рРНК и 72 — белок-кодирующие гены (**Рисунок 5**).

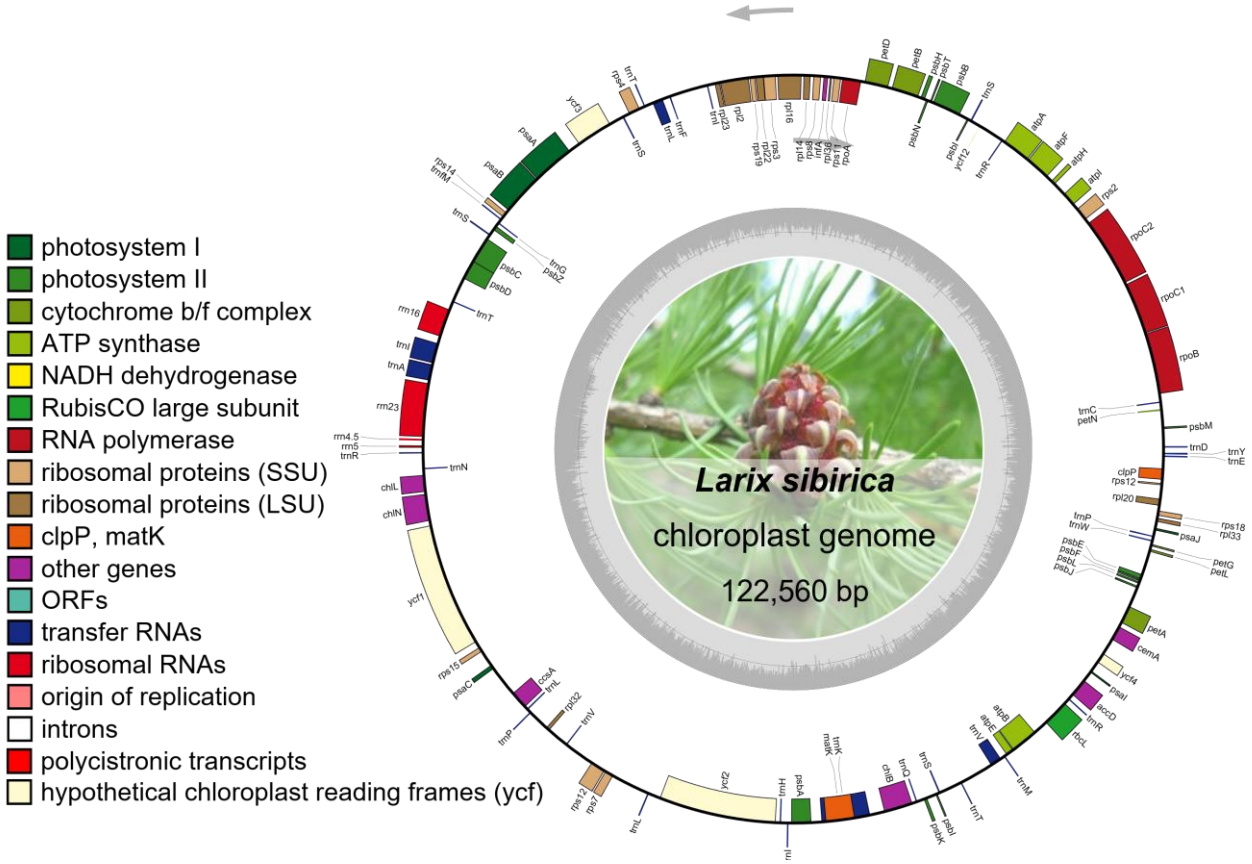

Рисунок 5. Карта расположения генов в хлоропластном геноме лиственницы сибирской. Гены, принадлежащие к разным функциональным группам, отмечены соответствующими цветами. Темно-серым и светло-серым цветами на внутреннем круге изображены GC- и AT-состав, соответственно.

Митохондриальный геном. На основании сборки коротких парно-концевых ридов (PE) Illumina из образцов, обогащенных митохондриальной ДНК (мтДНК), и mate-pair ридов (MP) Illumina из проекта полногеномного секвенирования лиственницы сибирской, а также после сопоставления гибридной сборки с длинными ридами с базой митохондриальных последовательностей растений, было собрано девять митохондриальных контигов общей длиной 11,7 млн. п.н. (максимальная длина контига 4008762 п.н.). На текущий момент это самый длинный митохондриальный геном из известных. Оценка правильности сборки с помощью REAPR показала долю безошибочных нуклеотидов 92,13%, что сопоставимо с 86% для референсного генома человека GRCh37 или 90% для генома *Caenorhabditis elegans* (Hunt et al. 2013). В ходе аннотации всего было идентифицировано 40 белок-кодирующих генов, 3 гена рРНК и 31 ген тРНК (**Рисунок 6**). С использованием комбинированной библиотеки повторов был идентифицирован 7691 повтор, что составляет 11% от 11,7 млн.п.н сборки. Причины чрезвычайно большого размера митогеномов у растений, сильно различающихся по длине у разных видов, до сих пор полностью не выяснены, но могут быть по крайней мере частично объяснены высокоизменчивым числом мобильных генетических элементов, пролиферацией ретротранспозонов, генерацией повторяющейся ДНК путем

рекомбинации и переносом чужеродных последовательностей из пластидной или ядерной ДНК или посредством горизонтального обмена мтДНК (Kan et al. 2020).

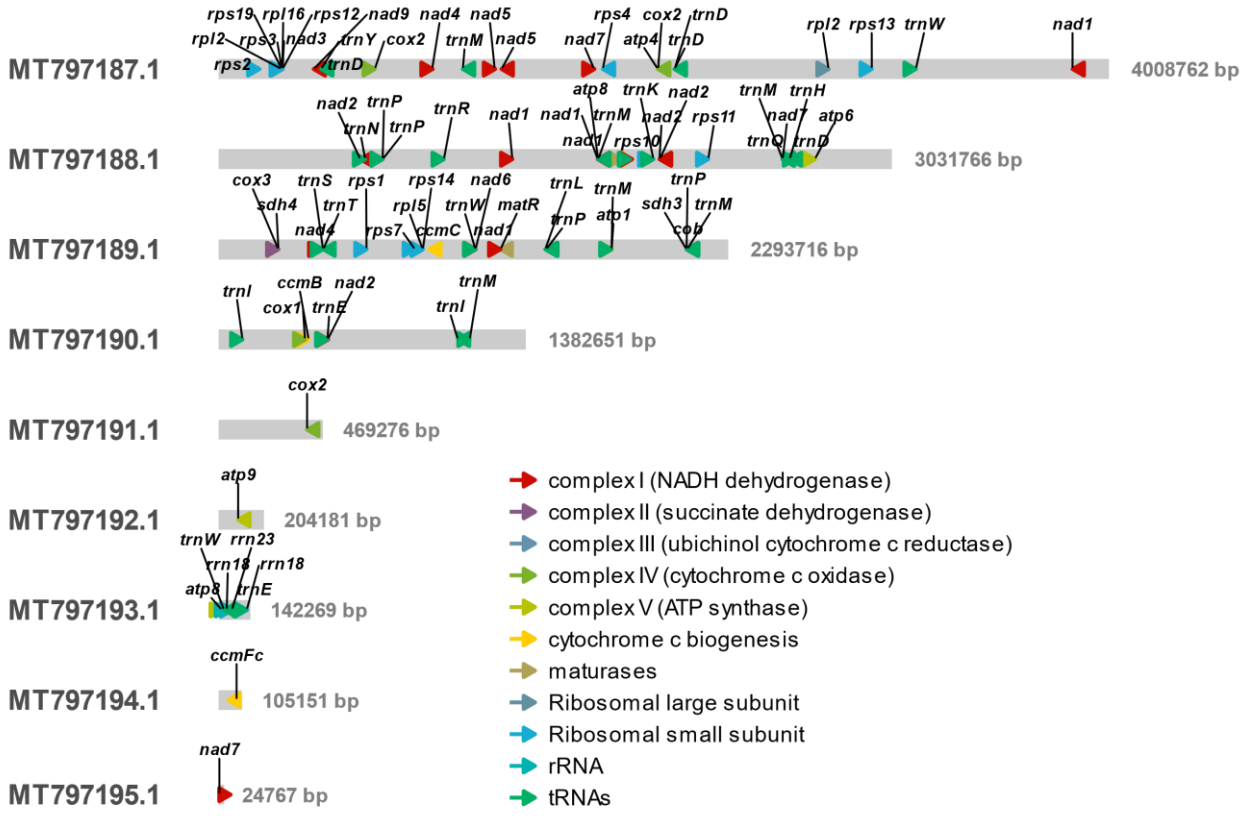

Рисунок 6 — Карта расположения генов в митохондриальном геноме лиственницы сибирской. 9 скаффолдов составляют 11,7 млн. п.н.

Предсказание сайтов начала транскрипции (TSS) в полногеномных сборках ели белой, ели норвежской, сосны ладанной, сосны сахарной и лиственницы сибирской. Использование TSSPlant позволило предсказать 62 420 позиции TSS у *L. sibirica*, 22 633 у *P. taeda*, 25 889 у *P. Abies* и 44 651 *у P. glauca*. Чтобы выбрать наиболее вероятную позицию TSS среди множества, предсказанных для данного гена, мы сравнили длину каждой 5'-UTR с распределением длин 5'-UTR у четырех видов растений: двух двудольных, *A. thaliana* и *Populus trichocarpa*, и двух однодольных растения *Oryza sativa* и *Sorghum bicolor*. При сравнении длин 5′-UTR, были выбраны такие предсказанные позиции TSS, которые давали длину 5'-UTR, наилучшим образом соответствующую их теоретическому распределению. После фильтрации предсказаний на предмет попадания в кодирующую область и выбора позиций с наибольшей вероятностью соответствия теоретическому распределению, 23 016 позиций были идентифицированы как предполагаемые TSS для *L. sibirica*, 10 367 для *P. abies*, 16 629 для *P. glauca* и 9 149 для *P. taeda*.

Все модели генов с соответствующими предсказанными TSS были размещены в геномном браузере Persephone и доступны по адресу https://web.persephonesoft.com/. 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 7**), что хорошо соответствует каноническому расположению TATA-бокса, поскольку от 30 до 50% эукариотических промоторов содержат TATA-бокс в положении на 15-40 п.н. выше TSS.

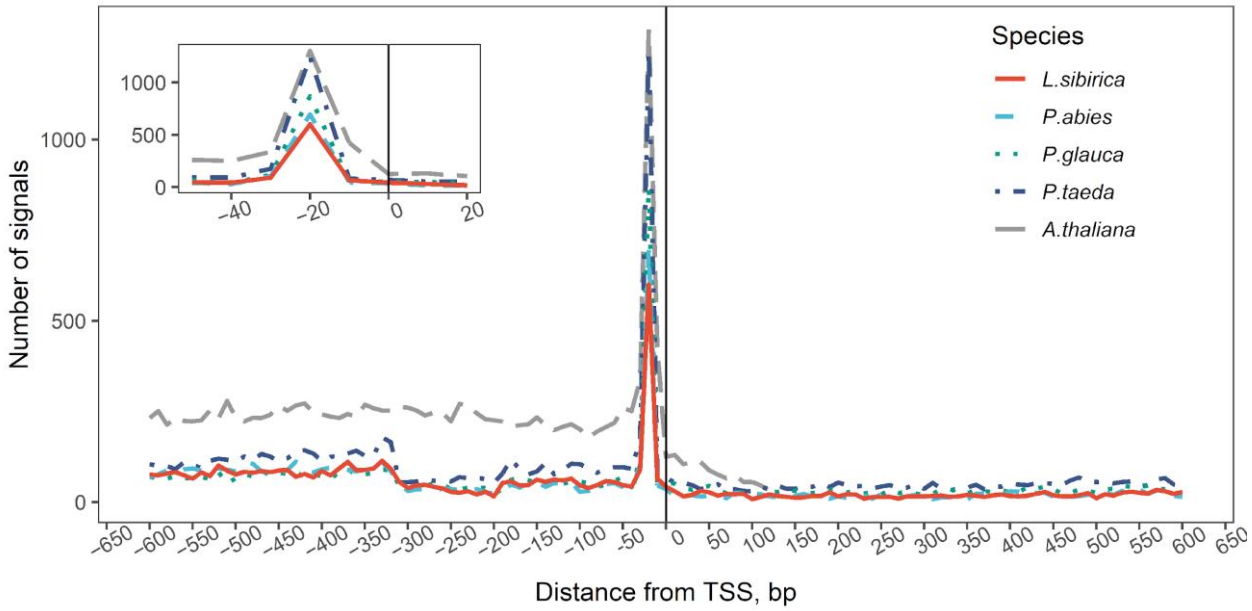

Рисунок 7. Частота мотива TATA(A/T)A(A/T) в TSS-центрированных промоторах для четырех видах хвойных и *A. thaliana*.

Изменение стандартной свободной энергии ДНК-дуплекса в последовательности генома является индикатором промоторной области и успешно применяется для предсказания промоторов. Данный подход был использован в качестве подтверждающего доказательства для промоторов, предсказанных TSSPlant. На консенсусных последовательностях промоторов профиль свободной энергии показывает пик около –40 п.н. и резкое снижение около предполагаемого TSS (**Рисунок 8**).

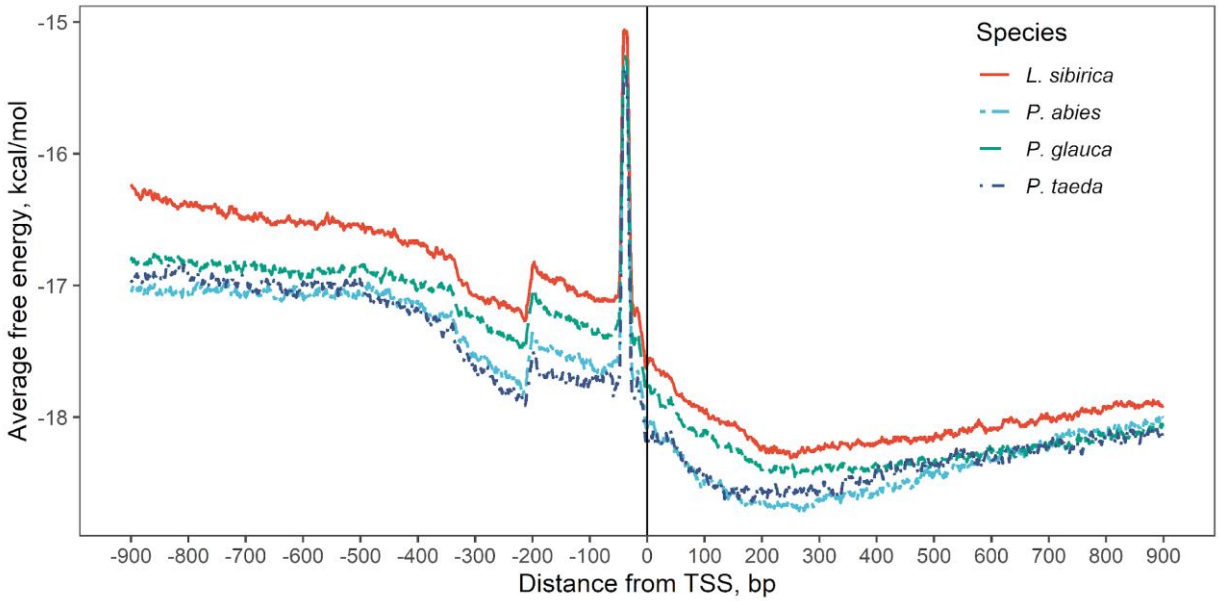

Рисунок 8. Распределение свободной энергии ДНК-дуплексов вокруг позиции TSS предсказанной TSSPlant.

Разработка микросателлитных маркеров для оценки генетического разнообразия лиственниц сибирской, Гмелина и Каяндера. В геномной сборке было найдено 1015 участков с высокоповторяющимися трех-, четырех-, пяти- и шестинуклеотидными мотивами. После отбора тандемных повторов по минимальному числу повторения мотивов и положению в контиге при помощи сервиса WebSat был выполнен дизайн праймеров для выбранных 222 локусов. Последовательности праймеров

были отфильтрованы, чтобы исключить повторы и попадание в органелльную ДНК. Набор из 60 пар праймеров проходил дальнейшее тестирование и подбор условий ПЦР. Только 23 локуса проявили устойчивую амплификацию в подобранных условиях. По результатам тестирования на выборках из 4 популяций было отобрано 14 полиморфных локусов (**Таблица 2**), пригодных для проведения популяционно-генетического анализа.

Таблица 2. SSR-локусы, полиморфные для трех видов лиственницы

Локус	Мотив	Размеры продукта	Число аллелей
Ls_752897	$(AAG)_{15}$	216-264	12
Ls_417667	$(AAT)_{15}$	207-243	5
Ls_954234	$(ATT)_{15}$	171-204	10
Ls_611965	$(CAG)_{15}$	222-276	7
Ls_1247092(2)	$(CTT)_{15}$	201-228	7
Ls_3765334	$(GAG)_{15}$	174-213	5
Ls_840190	$(TAC)_{15}$	216-249	7
Ls_1008427	$(ATAG)_{10}$	152-174	5
Ls_980491	$(CTAT)_{10}$	204-240	3
Ls_2552367	$(CTAT)_{10}$	184-196	4
Ls_3952800	$(TATG)_{10}$	200-264	9
Ls_2672894	$(TTTG)_{10}$	152-164	3
Ls_305132	$(GTCGGA)_{7}$	210-240	6
Ls_4040657	$(TCACTT)_{7}$	194-218	3

На основании аллельных частот 14 локусов были рассчитаны основные показатели генетической изменчивости исследованных популяций лиственниц сибирской, Гмелина и Каяндера. 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 (H_O = 0,558) и ожидаемой (H_E = 0,491) гетерозиготности гораздо большую изменчивость показывают выборки из популяций лиственницы сибирской. Такая картина вполне согласуется с литературными данными (Oreshkova, Belokon, and Jamiyansuren 2013) и может быть объяснена условиями, в которых произрастают данные виды.

Оценка популяционной структуры на основе *F*-статистик Райта показала, что индекс фиксации особи относительно популяции в среднем составляет около 8% (F_{IS} = -0,077), относительно вида 7% (F_{IT} = 0,074) (Guries and Ledig 1982). Приблизительно 13% от всей наблюдаемой изменчивости (F_{ST} = 0,137) 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 2**).

Уровень генетической дифференциации между исследованными популяциями был определён с использованием стандартного генетического расстояния (D_N) Нея; значения D_N между популяциями лиственницы варьируют в достаточно широких пределах: от 0,09 до 0,32. Анализ индивидуальных генотипов (генотипических дистанций) изученных видов лиственницы также показал подразделенность популяций и соответствие их географическому расположению: наиболее генетически удалёнными друг от друга оказались выборки лиственницы Каяндера из Якутии и лиственницы сибирской из Хакасии (D_N = 0,323), тогда как наименьшее значение генетического расстояния наблюдалось между выборками одного вида — лиственницы сибирской (D_N = 0,092). Установленный уровень дифференциации включенных в исследование выборок из популяций лиственницы наглядно показывает расположение популяций на плоскости двух координат (**Рисунок**).

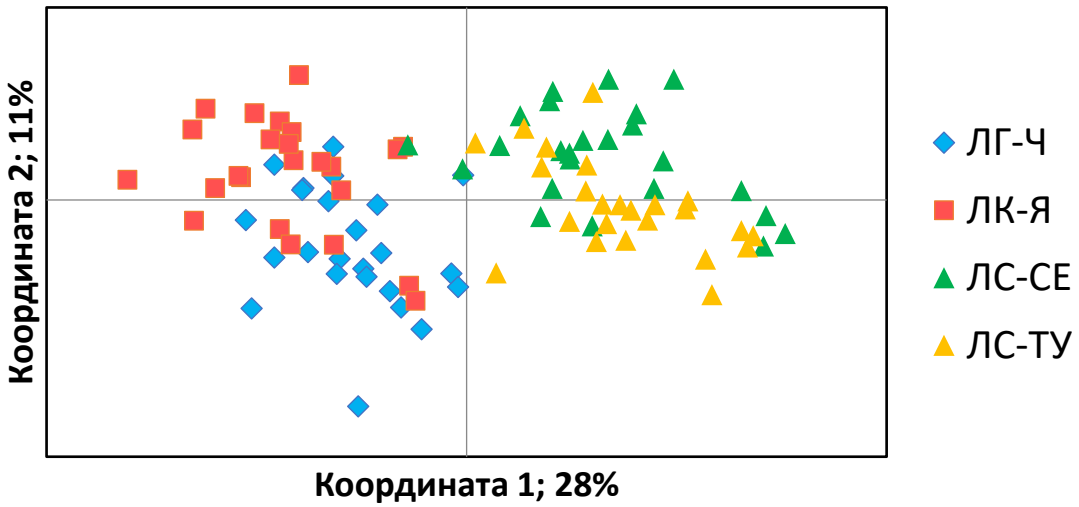

Рисунок 8. Проекция изученных выборок лиственницы на плоскости двух координат по данным анализа главных координат (PCoA) матрицы генотипических расстояний. ЛГ — лиственница Гмелина, ЛК — лиственница Каяндера, ЛС— лиственница сибирская.

Таким образом, разработанные локусы могут служить в качестве диагностических для генетической маркировки выборок из разных географических районов. Приведенные оценки межпопуляционной изменчивости согласуются с ранее опубликованными данными, как по анализу SSR-маркеров (Oreshkova, Belokon, and Jamiyansuren 2013), так и по анализу аллозимных маркеров для лиственниц сибирской и Гмелина (Абаимов 2009; Ларионова, Орешкова (Яхнева), and Абаимов 2004).

ЗАКЛЮЧЕНИЕ

Работа с полными геномами хвойных часто представляет собой нетривиальную задачу, так как из-за огромных размеров требуется серьезные вычислительные ресурсы (время вычислений, объем памяти), необходимые для обработки полногеномных данных. Структурная аннотация полногеномной сборки лиственницы сибирской на 448 ядрах вычислительного кластера заняла 22 дня, даже используя оптимизированную программу поэтапной сборки, 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 лиственницы сибирской, а также получены сборки транскриптомов нескольких тканей. Эти данные представляют собой первый публично доступный геномный ресурс для рода *Larix*. Сравнение полученных в ходе аннотации белок-кодирующих генов с набором генов *A. thaliana* 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*. 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 нескольких 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 части длинных прочтений Oxford Nanopore, по оценке RepeatMasker, составила 66% п.н., что близко к таковым оценкам для

лиственниц японской и Кемпфера, 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. Это соответствует эпохе плиоцена, когда климат Земли был значительно теплее, чем сегодня. В Сибири этот период характеризовался, вероятно, сравнительно более теплыми и влажными условиями, чем в настоящее время. 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 меньшее число повторов у лиственницы по сравнению с другими хвойными 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. Но сравнение с другими сборками показало, что меньшее число повторов у лиственницы по сравнению с другими хвойными вероятно объективно и не может быть объяснено только фрагментарностью сборки.

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 лиственницы сибирской. Общая длина митохондриального генома составила 11,7 млн. п.н., что на текущий момент является самым большим митохондриальным геномом из известных. Причины чрезвычайно большого размера митогеномов у растений, сильно различающегося у разных видов, до сих пор полностью не выяснены, но могут быть, по крайней мере, частично объяснены переменным и очень изменчивым числом мобильных генетических элементов, интронов и связанных с плазмидами последовательностей.
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 часть повторов была слишком фрагментирована, чтобы быть включенной в окончательную сборку, однако вероятно, что лиственницы действительно имеет меньше повторов, чем другие виды хвойных. Это коррелирует также с меньшим размером генома по сравнению с другими видами хвойных, что подтверждает роль и вклад повторяющихся мобильных элементов в размер генома. 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. Это соответствует эпохе плиоцена, когда климат Земли был значительно теплее, чем сегодня. В Сибири этот период характеризовался, вероятно, сравнительно более теплыми и влажными условиями, чем в настоящее время.
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 так же средне- и высоко-полиморфные спектры для лиственниц Гмелина и Каяндера. 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. Данные маркёры имеют большое практическое значения для паспортизации лесосеменных насаждений и клоновых плантаций, а также для ДНК-идентификации происхождения древесины и растительного материала в борьбе с нелегальными рубками и контроле за оборотом лесопосадочного материала.

Список работ, опубликованных по теме диссертации:

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.

Прочие публикации:

1. **Bondar, E. I**. Annotation of Siberian larch reference genome, the only seasonal senescence genus in Pinaceae / E. I. Bondar, S. I. Feranchuk, K. A. Miroshnikova, V.V. Sharov, D.A. Kuzmin, N. V. Oreshkova, K. V. Krutovsky // Высокопроизводительное секвенирование в геномике : Тезисы III Всероссийской конференции, Новосибирск, 19-24 июня 2022 г. – 2022. – С. 24.
2. **Bondar, E. I.** Annotation of Siberian Larch Genome Draft Assembly / E.I. Bondar, S.I. Feranchuk, V.V. Birukhov, D.A. Kuzmin, V.V. Sharov, N.V. Oreshkova, K.V. Krutovsky // Plant Genetics, Genomics, Bioinformatics, And Biotechnology : Abstracts of The 6th International Scientific Conference. – 2021. – P. 40.
3. **Bondar, E. I**. Genome-wide prediction of transcription start site in four conifer species / E. I. Bondar, V. V. Sharov, D. A. Kuzmin, T. V. Tatarinova, K. V. Krutovsky // Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2020). – 2020. – С. 300-301. DOI: 10.18699/BGRS/SB-2020-188/
4. Putintseva, Yu. A. Siberian larch (*Larix sibirica* Ledeb.) Mitochondrial genome, the largest currently known mitogenome / Yu.A. Putintseva, **E.I. Bondar**, V.V. Sharov, E.P. Simonov, N.V. Oreshkova, D.A. Kuzmin, Yu.M. Konstantinov, V.N. Shmakov, V.I. Belkov, M.G. Sadovsky, K.V. Krutovsky // Conservation of Forest Genetic Resources: Proceedings of the 6th International Conference. – Kokshetau publishing house «World of Printing», IE «Ustyugova». – 2019. – P. 19.
5. **Bondar, E. I.** Sequencing and assembly of mitochondrial genomes in three conifer species *Larix sibirica*, *Pinus sibirica* and *Pinus sylvestris* / E.I. Bondar, A. Kirichenko, V.V. Sharov, Yu.A. Putintseva, N.V. Oreshkova, S.I. Feranchuk, Yu.M. Konstantinov, V.N. Shmakov, V.I. Belkov, D.A. Kuzmin, M.G. Sadovsky, K.V. Krutovsky. // Systems biology BGRS/SB-2018 : The Eleventh International Conference on bioinformatics of genome regulation and structure. – 2018. – P. 153.
6. **Бондар, Е. И.** Разработка микросателлитных маркеров лиственницы сибирской (*Larix sibirica* Ledeb.) на основе полногеномного *de novo* секвенирования / Е. И. Бондар // Биология : Материалы 55-й Международной научной студенческой конференции. – 2017. – С. 156.

7. Sadovsky, M.G. *L. sibirica* Ledeb. chloroplast genome yields unusual seven-cluster structure / M.G. Sadovsky, Yu.A. Putintseva, **E.I. Bondar**, K.V. Krutovsky // IWBBIO 2016 : Proceedings Extended abstracts of International work-conference on Bioinformatics and biomedical engineering, Granada (Spain). – 2016. – P. 360.
8. **Bondar, E. I.** Siberian larch chloroplast genome analysis over triplet frequency distribution / E.I. Bondar, Yu.A. Putintseva, K.V. Krutovsky // Systems biology : Abstracts of the tenth international conference on bioinformatics of genome regulation and structure. – 2016. – P. 48.
9. **Bondar, E. I.** Comparative studying of multicluster structure of chloroplast genomes / E.I. Bondar, M.G. Sadovsky, Yu.A. Putintseva, M.Yu. Senashova // Systems Biology and Bioinformatics (SBB-2016) : Abstracts of The eighth international young scientists school. – 2016. – P. 12.
10. **Bondar, E. I.** Assembly and annotation of Siberian larch chloroplast genome and the search for single nucleotide polymorphisms / E.I. Bondar // Systems Biology and Bioinformatics : Abstracts of the 7th International Young Scientists School. – 2015. – C. 13.
11. **Бондар, Е. И.** Изучение хлоропластного генома лиственницы сибирской (*Larix sibirica* Ledeb.) и разработка полиморфных хлоропластных маркеров / Е.И. Бондар, Ю.А. Путинцева, Н.В. Орешкова, К.В. Крутовский // Материалы 4-го международного совещания “Сохранение лесных генетических ресурсов Сибири”. – 2015 г. – С. 20-21;
12. **Бондар, Е. И.** Сборка и аннотирование хлоропластного генома лиственницы сибирской (*Larix sibirica* Ledeb.) и поиск маркеров (SNPs) / Е.И. Бондар // Проспект Свободный-2015 : Сборник материалов Международной конференции студентов, аспирантов и молодых ученых, Красноярск, 15-25 апреля 2015 г. – 2015. – С. 6-7.
13. **Бондар, Е. И.** Сборка и аннотирование хлоропластного генома лиственницы сибирской / Е.И. Бондар // МНСК-2015 (Биология) : Материалы 53-й Международной научной студенческой конференции / Новосиб. гос. ун-т. – Новосибирск. – 2015. – С. 178.
14. **Бондар, Е. И.** Сборка и сравнительный анализ хлоропластных геномов хвойных рода *Larix* / Е.И. Бондар // Третья летняя школа по биоинформатике : Сборник тезисов. – 2015. – С. 9.