

Федеральное государственное бюджетное научное учреждение
«Федеральный исследовательский центр Институт цитологии и генетики
Сибирского отделения Российской академии наук» (ИЦиГ СО РАН)

На правах рукописи

Пронозин Артем Юрьевич

**Поиск, структурно-функциональная аннотация и анализ
разнообразия последовательностей
длинных некодирующих РНК в пан-транскриптоме
кукурузы *Zea mays* L.**

1.5.8. Математическая биология, биоинформатика

Диссертация на соискание учёной степени кандидата биологических наук

Научный руководитель:
д.б.н., г.н.с. ИЦиГ СО РАН,
доцент Афонников Д.А.

Новосибирск 2026

Оглавление	
Список использованных сокращений	5
Введение	6
Актуальность темы исследования	6
Цели и задачи работы	8
Научная новизна	9
Теоретическая и практическая ценность	10
Положения, выносимые на защиту	11
Апробация работы	11
Публикации	12
Структура и объем работы	12
Личный вклад автора	12
Благодарности	12
ГЛАВА 1. Обзор литературы	13
1.1 Общая характеристика днРНК	13
1.2 Классификация днРНК по положению в геноме	15
1.3 Молекулярные функции днРНК	17
1.3.1 Механизмы взаимодействия днРНК и микроРНК	22
1.3.2 днРНК в альтернативном сплайсинге	24
1.4 Экспрессия днРНК	28
1.5 Биологическая роль днРНК у растений	31
1.6 Эволюция днРНК	35
1.7 Методы секвенирования для исследования днРНК	37
1.8 Биоинформатические методы идентификации днРНК	40
1.9 Пан-геномы и пан-транскриптомы	44
1.9.1 Пан-геномы	44
1.9.2 Пан-транскриптомы	48
1.9.3 днРНК в пан-геномах и пан-транскриптомах	49
1.9.4 Биоинформатические методы анализа пан-геномов и пан-транскриптомов	50
1.10 Конвейеры обработки данных	54
1.11 Кукуруза и экономическая значимость	55

1.12 Заключение по обзору литературы и постановка задачи исследования.....	60
ГЛАВА 2 Материалы и методы.....	62
2.1 Разработка конвейера ICAnnoLncRNA для идентификации и анализа днРНК в транскриптомах растений.....	62
2.1.1 Предварительная обработка данных	64
2.1.2 Идентификация днРНК.....	67
2.1.3 Выявление видоспецифичности днРНК.....	71
2.1.4 Анализ структурных особенностей днРНК и их экспрессии.....	72
2.1.5 Платформа для реализации конвейера.....	74
2.1.6 Описание данных используемых для проверки работы биоинформатического конвейера ICAnnoLncRNA	74
2.2 Разработка конвейера РТА для анализа пан-транскриптома.....	75
2.2.1 Сборка транскриптомов	77
2.2.2 Сборка пан-транскриптома некодирующих последовательностей.....	77
2.2.3 Сборка пан-транскриптома кодирующих последовательностей.....	80
2.2.4 Анализ структуры пан-транскриптома.....	81
2.2.5 Анализ нуклеотидных вариаций последовательностей пан-транскриптомов.....	82
2.2.6 Анализ синонимических и несинонимических замен.....	83
2.2.7 Транскриптомные данные инбредных линий кукурузы.....	84
2.2.8 Проверка качества сборки транскриптомов.....	88
ГЛАВА 3 Применение конвейера ICAnnoLncRNA для анализа транскриптомов кукурузы и арабидопсиса.....	90
3.1 Результаты анализа транскриптомов кукурузы.....	90
3.1.1 Обучение и валидация модели предсказания днРНК.....	90
3.1.2 Идентификация днРНК.....	92
3.1.3 Классификация днРНК.....	94
3.1.4 Анализ структурных характеристик днРНК.....	95
3.1.5 Анализ видоспецифичности днРНК.....	98
3.1.6 Анализ экспрессии днРНК в разных тканях кукурузы.....	101
3.2 Проверка работы разработанного конвейера ICAnnoLncRNA на данных <i>A. thaliana</i>	103

3.3 Обсуждение.....	105
3.3.1 Сравнение характеристик разработанного конвейера ICAAnnoLncRNA с существующими.....	105
3.3.2 Оценка точности идентификации днРНК и сравнение их структурных характеристик с существующими данными.....	108
ГЛАВА 4 Сборка и анализ пан-транскриптома инбредных линий кукурузы	112
4.1 Результаты.....	112
4.1.1 Сборка пан-транскриптома линий кукурузы.....	112
4.1.2 Проверка качества сборки транскриптомов.....	114
4.1.3 Анализ структуры пан-транскриптомов.....	116
4.1.4 Анализ нуклеотидного разнообразия пан-транскриптомов.....	119
4.1.5 Анализ синонимических и несинонимических замен.....	122
4.1.6 Разнообразие днРНК, различающихся по расположению в геноме.....	124
4.2 Обсуждение.....	126
4.2.1 Пан-транскриптом инбредных линий кукурузы в сравнении с другими пан-геномами и пан-транскриптомами	126
Заключение	133
Выводы	136
Список литературы	139

Список использованных сокращений

ОНП - single nucleotide polymorphism, однонуклеотидные полиморфизмы

РНК - Рибонуклеиновая кислота

днРНК - Длинные некодирующие РНК

мРНК - Матричная РНК

ДНК - Дезоксирибонуклеиновая кислота

RNA-seq - RNA sequencing, массовое высокопроизводительное секвенирование РНК

ОРС - открытые рамки считывания

FPKM - Fragments Per Kilobase of transcript per Million mapped reads, количество фрагментов для конкретного экзона на длину гена на миллион картированных прочтений

TPM - Transcript per Million mapped reads, количество транскриптов на длину гена на миллион картированных прочтений

ФС - фактор сплайсинга

CDS - кодирующая последовательность РНК

Введение

Актуальность темы исследования

Длинные некодирующие РНК (днРНК) представляют собой класс линейных или кольцевых молекул РНК длиной от 200 нуклеотидов, некодирующих белок. Подобно мРНК, днРНК содержат сайты сплайсинга, последовательности экзонов (фрагментов зрелой днРНК) и поли(А)-хвост (Ulitsky, 2013). Регуляция их экспрессии осуществляется аналогично мРНК, однако показано, что их экспрессия более тканеспецифична, а ее уровень ниже, чем у мРНК (Derrien et al., 2012). Долгое время их функции были слабо изучены, и эти молекулы рассматривались как транскрипционный шум (Yang et al., 2022). Дальнейший анализ показал, что днРНК участвуют в большом количестве важнейших процессов у животных и растений: регуляции экспрессии генов (Karlik, 2019), формировании структуры макромолекулярных комплексов (Tsai et al., 2010), взаимодействуют с белками (Sousa et al., 2001), обуславливают развитие патогенеза у животных и человека (Kazemzadeh, 2015).

По сравнению с животными функции днРНК у растений изучены менее детально. Однако экспериментально показано, что они участвуют в регуляции устойчивости к холодовому, солевому и тепловому стрессу (Zhu и Wang, 2012), влияют на устойчивость к гипоксии (Cheng et al., 2012), принимают участие в развитии плодов, корней и листьев (Huang et al., 2018) и в большом количестве других процессов роста и развития растений, ответа на стресс (Traubenik, 2024). днРНК выполняют функции как генетических, так и эпигенетических регуляторов экспрессии генов, могут действовать как цис-регуляторные элементы, взаимодействуя с соседними генами, или как транс-регуляторные элементы, оказывая влияние на гены, расположенные на удалении от места их синтеза (Wang et al., 2023). Это разнообразие функций делает днРНК важным объектом для понимания молекулярных механизмов роста и развития растений, их ответа на биотический и абиотический стресс. Однако до сих пор только для небольшой части днРНК их функциональная роль была охарактеризована экспериментально. Для большинства днРНК функции остаются неизвестными.

Важную информацию о функциональной роли днРНК может предоставить сравнительный анализ последовательностей из геномов различных организмов. По современным представлениям, функционально важные последовательности являются более консервативными. Анализ эволюции днРНК у животных (Hezroni et al., 2015) и растений (Hao et al., 2015) продемонстрировал ряд важных особенностей этих молекул. Число идентифицированных днРНК у вида растений коррелирует с размером генома, и наибольшее их количество среди 26 видов оказалось у кукурузы (Chen и Zhu, 2022).

В то время как структурные свойства днРНК (распределение по длинам, числу экзонов, уровню экспрессии относительно мРНК) у разных видов растений достаточно схожи (Chen и Zhu, 2022), число общих днРНК выше среди родственных видов (в пределах одного рода), при этом их доля резко убывает по мере увеличения эволюционного расстояния (Sang et al., 2021). Таким образом, более 70% днРНК является видоспецифичными (Hezroni et al., 2015). Во многих работах показана высокая внутривидовая вариабельность первичной структуры днРНК по сравнению с мРНК (Wang et al., 2015). Они демонстрируют большую долю однонуклеотидных замен по сравнению с мРНК (Chen и Zhu, 2022).

В последнее время в связи с прогрессом в методах секвенирования и сборки геномов стало возможным реконструировать не один, а множество геномов одного вида. Это позволяет детально изучать структурные различия геномов одного таксона, такие как вариации наличия-отсутствия генов (PAV). Такие вариации принято описывать в рамках концепции пан-генома, под которым подразумевают неизбыточный набор генов, представленных в геномах таксона, как правило, вида (Rasko et al., 2008). Пан-геномы бывают открытыми (при добавлении новых представителей вида количество новых генов неограниченно растет) и закрытыми (для которых существует верхний предел количества генов). Аппроксимацией пан-генома может служить пан-транскриптом, определенный на основе генов, экспрессирующихся у разных представителей таксона. В этом случае рассматриваются вариации, оцененные на основе экспрессии генов (ePAV), при этом структурные особенности пан-генома и пан-транскриптома оказываются близкими (Hirsch et al., 2014).

Для сельскохозяйственных растений модель пан-генома активно используется при поиске новых белок кодирующих генов, оценке их разнообразия, структурных особенностей и эволюции. Она позволяет разделить гены на две группы которые присутствуют у всех представителей вида (ядро) и гены, которые присутствуют лишь у некоторых представителей вида (оболочка и облако). Гены второй группы обычно связаны с адаптацией растений к условиям окружающей среды и устойчивостью к болезням. Поэтому они вызывают большой интерес исследователей. Следует отметить, что для описания разнообразия длинных некодирующих РНК модель пан-транскриптома ранее не применялась.

У культурных растений одним из наиболее активно изучаемых является геном кукурузы *Zea mays*. Кукуруза является одной из ведущих зерновых культур мирового земледелия (Shindin, 2009). Широкому распространению и увеличению производства кукурузы способствуют её высокая урожайность и широкие возможности использования как продукта питания, так и ценного корма для сельскохозяйственных животных (Korobka, et al., 2015). Геном кукурузы является одним из наиболее полно секвенированных и аннотированных. Для кукурузы уже проведены исследования по сборке пан-геномов и пан-транскриптомов, однако только для белок кодирующих генов. В связи с этим представляет интерес изучение структуры и функций днРНК кукурузы на основе имеющихся данных по секвенированию пан-транскриптома.

Цели и задачи работы

Поиск длинных некодирующих РНК в пан-транскриптоме кукурузы, оценка их структурного разнообразия и выявление закономерностей их эволюции на основе использования биоинформатических конвейеров.

Задачи, поставленные для достижения указанной цели, включали:

1. Создание биоинформатического конвейера для предсказания и аннотации последовательностей днРНК сельскохозяйственных растений на основе широкомасштабного анализа транскриптомов, его апробация на примере кукурузы и арабидопсиса.

2. Создание биоинформатического конвейера для сборки и анализа пан-транскриптомов сельскохозяйственных растений.
3. Сборка пан-транскриптомов днРНК и белок кодирующих генов инбредных линий кукурузы, сравнительный анализ их структурных особенностей.
4. Сравнительный анализ темпов накопления замен последовательностей мРНК и днРНК в пан-транскриптоме линий кукурузы.
5. Сравнительный анализ разнообразия антисмысловых, интронных и межгенных днРНК в масштабах пан-транскриптома линий кукурузы и скоростей их нуклеотидных замен.

Научная новизна

Впервые с использованием данных RNA-seq 503 инбредных линий кукурузы реконструированы пан-транскриптомы днРНК и мРНК и проведен их комплексный биоинформатический анализ. Показано, что оба пан-транскриптома являются закрытыми, при этом доля последовательностей, которые являются специфичными для транскриптомов отдельных линий кукурузы у днРНК выше, по сравнению с мРНК.

Впервые проведено сравнительное исследование эволюционных характеристик последовательностей днРНК и белок кодирующих генов кукурузы в составе пан-транскриптома. Установлено, что днРНК демонстрируют более высокое генетическое разнообразие и темп накопления нуклеотидных замен по сравнению с мРНК.

Выявлены эволюционные различия между антисмысловыми, межгенными и интронными днРНК, связанные с возможными различиями функций этих типов днРНК. Показано, что антисмысловые днРНК являются наиболее консервативными как по темпам накопления нуклеотидных замен, так и по составу последовательностей в геномах инбредных линий кукурузы; интронные днРНК характеризуются наибольшей скоростью мутаций, а межгенные днРНК демонстрируют максимальное разнообразие состава и его высокую специфичность у инбредных линий кукурузы.

Теоретическая и практическая ценность

Разработан биоинформатический конвейер ICAnnoLncRNA идентификации и анализа днРНК в транскриптомных данных на основе методов машинного обучения с возможностью настройки параметров, специфических для референсных геномов растений разных видов. Конвейер обеспечивает не только предсказание днРНК, но и их классификацию на основе геномной локализации, анализ экспрессии и тканеспецифичности. Предложенные опции повышают точность идентификации днРНК у разных видов растений, что было продемонстрировано при анализе транскриптомов кукурузы и арабидопсиса. Вычислительный конвейер ICAnnoLncRNA свободно доступен по адресу: <https://github.com/artempronozin95/ICAnnoLncRNA-identification-classification-and-annotation-of-LncRNA>

Создан автоматизированный конвейер РТА (Pan-Transcriptome Assembler) для сборки и анализа кодирующих и некодирующих РНК, позволяющий исследовать особенности эволюции мРНК и днРНК в рамках модели пан-транскриптома. Вычислительный конвейер РТА свободно доступен по адресу:

<https://github.com/artempronozin95/PTA-Pan-Transcriptome-Assembler>

С помощью применения конвейеров выявлены основные закономерности пан-транскриптома днРНК инбредных линий кукурузы (большой по размеру в сравнении с мРНК размер пан-транскриптома, в котором преобладают гены облака); высокая консервативность антисмысловых днРНК; большая доля межгенных днРНК относится к облаку в сравнении с антисмысловыми и интронными. Эти закономерности могут быть использованы для детального изучения разнообразия днРНК у других видов растений.

Разработанные биоинформатические конвейеры могут быть адаптированы для анализа транскриптомов других сельскохозяйственных культур для идентификации днРНК в транскриптомных данных большого объема. Результаты выявления и аннотации днРНК с помощью разработанных конвейеров позволяют определить различия их структур, уточнить функциональную роль в молекулярных процессах растений.

Положения, выносимые на защиту

1. Биоинформатический конвейер ICAnnoLncRNA за счет применения алгоритмов машинного обучения и многоуровневой фильтрации артефактов обеспечивает высокую точность массового предсказания и аннотации длинных некодирующих РНК (днРНК) в транскриптомах растений.
2. Пан-транскриптомы мРНК и днРНК инбредных линий кукурузы имеют «закрытый» тип (ограничены по числу экспрессирующихся генов), при этом их структуры отличаются друг от друга – в пан-транскриптоме мРНК преобладают гены «ядра» и «оболочки», которые считаются эволюционно более древними, тогда как в пан-транскриптоме днРНК преобладают гены «облака», что свидетельствует об их относительно молодом эволюционном возрасте.
3. днРНК, различающиеся по положению в геноме, отличаются по темпам эволюции – межгенные и интронные днРНК имеют наибольшую скорость мутаций, тогда как антисмысловые днРНК являются наиболее консервативными по темпам нуклеотидных замен и по составу последовательностей.

Апробация работы

Основные результаты, полученные в рамках настоящей работы, были представлены на международных конференциях «Биоинформатика геномной регуляции и структурной системной биологии» (Новосибирск, 2022 - 2024 гг.); PlantGen (Новосибирск 2021-2025); Марчуковских научных чтениях (Новосибирск, 2021); Московская конференция по вычислительной молекулярной биологии (MCCMB) (Москва, 2021); Conference on Cereal Biotechnology and Breeding (CBB6) (Венгрия 2021). А также на съезде Вавиловского общества генетиков и селекционеров в 2024 году. Разработанные конвейеры запатентованы.

Публикации

По материалам диссертации опубликовано 5 статей в журналах из Перечня ВАК (все индексируются в РИНЦ, Scopus и Web of Science), в том числе 2 в зарубежных журналах, а также 10 тезисов в сборниках материалов научных конференций. Все работы опубликованы в соавторстве с коллегами.

Структура и объем работы

Диссертация состоит из девяти разделов: введение, обзор литературы, материалы и методы, двух разделов с результатами (разработка метода предсказания и аннотации днРНК и анализ пан-транскриптома), заключение, выводы, список литературы (348 источника), а также включает оглавление, приложение и список сокращений. Работа изложена на 170 страницах (включая Приложение), содержит 33 рисунка и 10 таблиц.

Личный вклад автора

Автором разработаны два биоинформатических конвейера: 1) для предсказания и аннотации днРНК, 2) для автоматической сборки и анализа пан-транскриптома. Также автор провел комплексный анализ выявленных днРНК и белок кодирующих генов в пан-транскриптом кукурузы.

Благодарности

Автор выражает искреннюю благодарность д.б.н. Афонникову Д.А. за научное руководство данной работой. Рассказову Д.А за предоставление доступа к ресурсам Центра коллективного пользования ИЦиГ СО РАН «Биоинформатика» для проведения расчетов. Землянской Е.В. и Ершову Н.И. за полезные советы по улучшению работы.

ГЛАВА 1. Обзор литературы

1.1 Общая характеристика днРНК

Некодирующие РНК (нкРНК) выполняют в геномах растений ряд важнейших функций, связанных с регуляцией экспрессии генов, гомеостазом физиологических параметров растений. Знание функциональных особенностей нкРНК позволяет целенаправленно улучшать различные свойства растений (Zhu и Wang, 2012; Wu et al., 2013; Deng et al., 2018), что обуславливает интерес к изучению этих молекул. Некодирующие РНК можно разделить на две группы: структурные некодирующие (рибосомные РНК, транспортные РНК, малые ядерные РНК, малые ядрышковые РНК) и регуляторные некодирующие (малые некодирующие РНК и длинные некодирующие РНК) (Назипова, 2021). Наименее изученными из всех перечисленных классов молекул являются длинные нкРНК (Kim и Sung, 2012; Joshi и Romanowska, 2020).

Длинные некодирующие РНК (днРНК) являются важным классом некодирующих РНК в живых организмах и представляют собой линейные или кольцевые молекулы РНК, содержащие более 200 нуклеотидов без открытых рамок считывания (Kim и Sung, 2012; Назипова, 2021). Функции и структура (как первичная, так и вторичная) днРНК слабо изучены в сравнении с белок кодирующими генами и генами малых некодирующих РНК. Однако участие днРНК выявлено в регуляции экспрессии генов (Karlik, 2019), формировании структуры макромолекулярных комплексов (Tsai et al., 2010), во взаимодействии с белками (Sousa et al., 2001), в патогенезе (Medvedeva et al., 2015). днРНК наиболее активно изучаются экспериментально и биоинформатически у человека и других модельных видов (мышь, крыса, дрозофила и т.д.). Эти исследования позволяют идентифицировать последовательности днРНК в геномах и транскриптомах, расшифровать важные особенности их структуры, регуляции экспрессии и функционирования, классифицировать их по структуре и функции. На сегодняшний день эти методы позволили идентифицировать большое количество последовательностей днРНК у животных. Так, согласно базе данных Gencode (Frankish et al., 2021), в геноме мыши описано около 13 000 генов днРНК. В геноме человека насчитывается более 18 000 генов днРНК (Гордиюк, 2014).

Изучение днРНК у животных позволило классифицировать их по внутриклеточной локализации (промоторные, энхансерные, теломерные, сигнальные и др.) (Li et al., 2014; Zhao et al., 2018) и по положению в геноме (межгенные, интронные, антисмысловые и др.) (Ma, 2013; De Quattro, 2017; Karlik, 2019; Chen, 2020). Эти исследования приводят к созданию специализированных баз данных днРНК, содержащих последовательности десятков тысяч транскриптов, а также информацию об их первичной структуре, экзон-интронной организации, геномном контексте и, в некоторых случаях, вторичной структуре (Griffiths-Jones, 2006; Amaral и Mattick, 2008). Вследствие отсутствия у первичной структуры днРНК четко выраженных функциональных сегментов (как, например, домены у белковых последовательностей), на сегодняшний день предсказание функции днРНК на основе первичной структуры является сложной задачей (Blythe, 2016; Бейлерли и Гареев, 2020; Ahmad et al., 2021).

Рост интереса к днРНК привел к тому, что экспериментальных методов стало недостаточно для исследования данных молекул. Поэтому для изучения этого класса РНК активно используются методы биоинформатики. В частности, они направлены на широкомасштабное распознавание последовательностей днРНК в геномах и транскриптомах (Zhao et al., 2018; Bryzghalov, 2021; Talyan et al., 2021), предсказание их взаимодействий с белками (Campalans, 2004), с мРНК и другими некодирующими РНК (Unver и Tombuloglu, 2020a), на изучение влияния днРНК на экспрессию других генов (Karlik, 2019), изучение механизмов их эволюции (Duret et al., 2006; Khorkova, 2015) и предсказание их вторичной структуры и функций (Graf и Kretz, 2020).

днРНК также известны и у растений. Интересно отметить, что , несмотря на ряд специфических особенностей (способность транскрибироваться РНК-полимеразой IV и V у растений), функциональные и структурные характеристики днРНК животных и растений имеют много общего (Golicz, 2018). Известно, что у растений днРНК принимают участие в регуляции устойчивости к холодовому, солевому и тепловому стрессу (Zhu и Wang, 2012; Wu et al., 2013; Deng et al., 2018), оказывают влияние на устойчивость к гипоксии (Cheng et al.,

2012), принимают участие в развитии плодов, корней и листьев (Ma, 2013; Huang et al., 2018). Однако, несмотря на растущее число исследований, посвященных структурно-функциональному анализу днРНК, данный тип молекул по-прежнему остается малоизученным. Это связано со множеством факторов, которые нужно учитывать при идентификации днРНК. Размер днРНК схож с размером некоторых белок кодирующих генов, что приводит к ошибкам распознавания. Также исследования показывают, что последовательности днРНК претерпевают быструю эволюцию, закономерности которой пока непонятны. Как правило, в процессе эволюции эти последовательности очень быстро накапливают замены так, что даже с последовательностями из близкородственных видов организмов они имеют слабую гомологию (Ponting, 2009; Derrien et al., 2012).

1.2 Классификация днРНК по положению в геноме

днРНК классифицируют по многим параметрам: длина транскрипта, расположение в геноме, взаимодействие с ДНК, РНК или белками, участие в процессах регуляции экспрессии генов (Karlik, 2019). Однако наиболее часто используют классификацию по положению в геноме относительно известных белок кодирующих генов. Согласно этой классификации, днРНК делятся на четыре крупных класса (Ponting, Oliver и Reik, 2009) рисунок 1.1:

- интронные — перекрываются с интроном гена.
- антисмысловые — ориентированы против направления транскрипции гена, кодирующего белок.
- межгенные — расположены между двумя локусами генов.
- смысловые — ориентированы в направлении транскрипции гена, кодирующего белок.

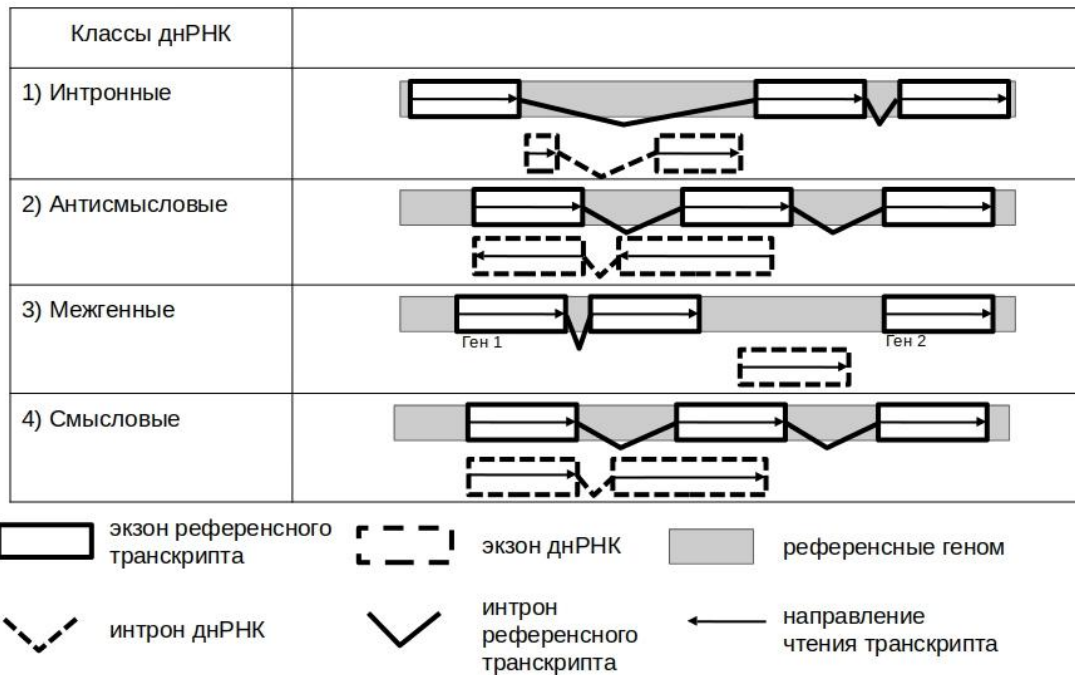


Рисунок 1.1. Классификация днРНК на основе расположения и ориентации относительно ближайших или перекрывающихся белок кодирующих генов.

Инициация транскрипции интронных днРНК происходит в направлении транскрипции гена-мишени, терминация происходит в области перекрывания с экзоном гена-мишени (Budak, Kaya и Cagirci, 2020).

Транскрипция антисмысловых днРНК начинается на 3'-конце гена, кодирующего белок. Эти днРНК разделяют на две группы в зависимости от цис- и транс-геометрической изомерии. Транскрипция цис-изомеров антисмысловой днРНК (цис-днРНК) происходит с противоположного конца локуса белок кодирующего гена, с которым данный транскрипт перекрывается. За счет этого днРНК имеет высокий уровень комплементарности с данным геном. Транскрипция транс-изомеров антисмысловой днРНК происходит с локуса соседнего белок кодирующего гена, с которым перекрывается данный транскрипт. Таким образом, подобные транскрипты могут перекрываться с несколькими генами, но при этом имеют низкий уровень комплементарности с ними (Britto-Kido et al., 2013).

Длина межгенных днРНК превышает 200 нуклеотидов, данные транскрипты не перекрываются с кодирующими белок генами. Особенности

транскриптов данного класса: высокая вариабельность, выраженная тканеспецифичность. В работе Kapusta et al. (2013) биоинформатическими методами было показано взаимодействие межгенных днРНК с мобильными элементами. Проведенное исследование предоставляет систематическую оценку вклада мобильных элементов в состав, эволюционное происхождение и регуляцию днРНК.

Смысловые днРНК транскрибируются в направлении транскрипции белок кодирующих генов. Это делает их выделение в отдельный класс днРНК или даже их идентификацию в качестве днРНК достаточно противоречивым. Структурные характеристики таких РНК (длина, количество экзонов и интронов) могут совпадать со структурой белок кодирующего гена. В связи с этим определяющим критерием при выявлении подобных транскриптов становится наличие или отсутствие кодирующего потенциала (Ma, 2013). Например, в работе (Ma, 2013) предполагается, что транскрипты данного класса днРНК могут полностью перекрываться с генами, кодирующими белок. Однако в работе (Budak, 2020) описываются примеры днРНК, возникших за счет альтернативного сплайсинга или укорачивания первого либо последнего экзона генов. Авторы предполагают, что эти гены могут представлять изоформы белков, кодируемых этими генами, а не днРНК. Например, в работах (Derrien et al., 2012, Budak, 2020) РНК, перекрывающиеся с экзоном гена, кодирующего белок, не считаются днРНК. Для растений таких транскриптов найдено мало, и ни для одного из них нет функциональной аннотации (Budak, 2020).

1.3 Молекулярные функции днРНК

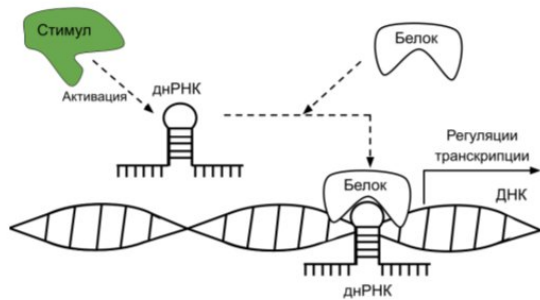
днРНК имеют несколько общих характеристик с генами, кодирующими белки, что затрудняет их идентификацию. Так, например, большинство днРНК транскрибируется с помощью РНК-полимеразы II. Они имеют сходные эпигенетические профили, наличие сигналов сплайсинга и полиаденилирования, а также сходные размеры экзонов и интронов (Ulitsky и Bartel, 2013; Лукина et al., 2022). Кроме того, некоторые днРНК имеют длину, превышающую 1000 нк. Таким образом, они могут содержать открытые рамки считывания (ОРС).

Например, в работе (Dinger et al., 2008) упоминается, что днРНК в среднем могут содержать несколько транслируемых ОРС, имеющих длину в пределах 50 кодонов. Однако наличие ОРС у некоторых днРНК не подразумевает наличие кодирующего потенциала. В большинстве случаев подобные ОРС либо не транслируются, либо приводят лишь к синтезу коротких пептидов без признаков функционального белка (Ransohoff, 2018).

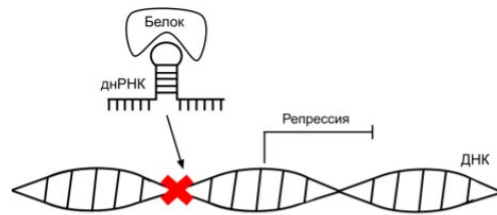
Отличительной особенностью днРНК является способность транскрибироваться РНК-полимеразой IV и V, особенно у растений (Xie et al., 2014). Большинство днРНК локализуются как в клеточном ядре, так и в цитоплазме (Pal и Rao, 2017). Также днРНК демонстрируют более низкий уровень экспрессии, при этом экспрессируются более специфично для определенных тканей (Татосян et al., 2020). Сравнение днРНК различных видов растений показало, что доля сходных последовательностей у разных видов мала. Однако днРНК одного вида демонстрируют достаточно высокий уровень консервативности первичной, вторичной и третичной структуры (Ahmad et al., 2021).

В работах (Li et al., 2014; Ahmad et al., 2021; Zhao et al., 2022) указывается, что на функционирование днРНК влияет их внутриклеточная локализация. В работах (Kopp и Mendell, 2018; Назипова, 2021; Ahmad et al., 2021; Zhao et al., 2022) выделяют пять основных функций днРНК, представленных на рисунке 1.2.

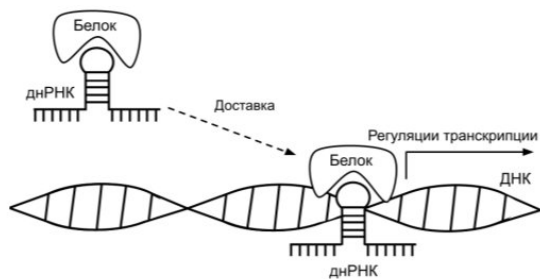
А Сигнальные



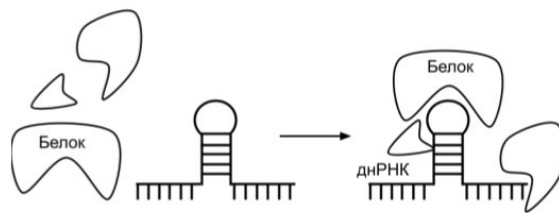
Б Ловушки



В Проводники



Г Каркасы



Д Энхансеры

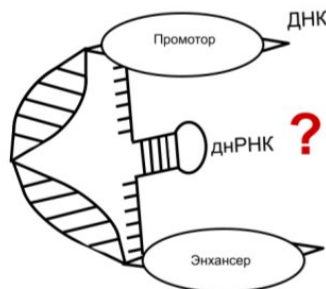


Рисунок 1.2. Молекулярные функции днРНК: А) Сигнальная: днРНК получает сигнал для взаимодействия с ферментами, модифицирующими хроматин, для регуляции транскрипции. Б) Ловушка: связывает белки и препятствует их присоединению к генам-мишеням. В) Проводник: направляет транскрипционные факторы к местам их функционирования, содействует необходимому состоянию хроматина. Г) Каркас: объединяет белки в комплексы и инициирует разные биологические процессы. Д) Энхансер: служит способом поддержания открытого состояния хроматина в области энхансеров.

Сигнальные днРНК

днРНК чаще всего экспрессируются в определенных тканях. В некоторых работах отмечается, что днРНК экспрессируются специфически в ответ на различные стимулы (Huarte и Rinn, 2010; Loewer et al., 2010; Hung et al., 2011; Назипова, 2021). В связи с этим можно предположить, что экспрессия некоторых днРНК находится под контролем транскрипционных факторов (Wang и Chang, 2011), осуществляется в определенный момент времени и в определенной ткани и может служить сигналом для запуска регуляторных процессов в клетке (Wang и Chang, 2011). Такая специфическая экспрессия характерна для днРНК, которые называют сигнальными (рисунок 1.3А). Сигнальные днРНК могут выступать в качестве элементов регуляции. Известно, например, что сигнальные днРНК взаимодействуют с модифицирующими хроматин ферментами, такими как гистоновые метилтрансферазы, чтобы подавить экспрессию их генов-мишеней посредством образования гетерохроматина (Chen, Wang и Yao, 2016). Такие взаимодействия и являются частью регуляторных и сигнальных функций данного типа днРНК.

днРНК приманки или ловушки

днРНК приманки связывают белки и тем самым препятствуют их взаимодействию с регуляторными районами генов-мишеней (Wang и Chang, 2011) рисунок 1.2Б. Эти днРНК косвенным образом репрессируют транскрипцию, блокируя микроРНК, модифицирующие комплексы (субъединицы хроматина), каталитические белки и факторы транскрипции (Wang и Chang, 2011; Kotake et al., 2011). Также днРНК выступают в качестве ловушки для микроРНК; данный механизм взаимодействия подробно описан в подразделе 1.3.1 «Механизмы взаимодействия днРНК и микроРНК».

днРНК проводники

днРНК также выполняют функцию группировки/локализации факторов транскрипции в определенных геномных локусах. Такие днРНК называют направляющими (рисунок 1.2В). днРНК направляют модификаторы хроматина и белковые комплексы или факторы транскрипции к определенным генам-

мишеням. Они способствуют локализации регуляторных молекул в определенных локусах для осуществления транскрипционной регуляции (Chen, Wang и Yao, 2016). днРНК имеют цис- и транс-геометрическую изомерию. Цис-изомеры направляют изменения хроматина, выступая в качестве комплементарных мишеней для малых РНК (таких как микроРНК, малые интерферирующие РНК) (Wang и Chang, 2011). Транс-изомеры доставляют ДНК-мишень посредством РНК-ДНК-взаимодействий (Ye et al., 2022).

днРНК каркасы

днРНК каркасы (скаффолды) играют важную роль при сборке макромолекулярных белковых комплексов и инициации различных биологических процессов (Yang, Froberg и Lee, 2014) рисунок 1.2Г. Подобные днРНК способны соединять несколько белков в один рибонуклеиновый комплекс, который позволяет подавлять или активировать транскрипцию в зависимости от наличия и природы вовлеченных днРНК и белков (Wang и Chang, 2011).

днРНК энхансеры

днРНК, транскрипция которых осуществляется с энхансерных локусов, называются днРНК энхансерами (рисунок 1.2Д). Как и все днРНК, их последовательности считываются в одном направлении, имеют поли(А)-хвост и сплайсируются (Lam et al., 2013). Их следует отличать от энхансерных РНК (эРНК), которые являются продуктом двунаправленной транскрипции, не сплайсированы и не имеют поли(А)-хвоста. Энхансерные днРНК на сегодняшний день довольно плохо изучены. Однако согласно гипотезе (Kim, Nenberg и Gray, 2015) синтез энхансерных днРНК может служить способом поддержания открытого состояния хроматина в области энхансеров для облегчения их взаимодействия с промоторами посредством образования петель. Некоторые исследования показывают вовлеченность энхансерных днРНК в процесс активации экспрессии генов-мишеней (Kim et al., 2010; Lai et al., 2013). Также в нескольких работах показано понижение экспрессии генов-мишеней при нокадауне ближайшей к ним энхансерной днРНК (Lai et al., 2013; Melo et al., 2013).

1.3.1 Механизмы взаимодействия днРНК и микроРНК

Особое внимание стоит уделить взаимодействию днРНК и микроРНК. В работе (Zhang et al., 2020) указывается на участие данных взаимодействий в регуляции устойчивости к болезням, холодовому, тепловому, биотическому и абиотическому стрессу у растений.

МикроРНК - это малые некодирующие РНК длиной 18-25 нт., транскрибируемые из интронов или межгенных регионов (Chen и Rajewsky, 2007). На сегодняшний день выделяют четыре типа взаимодействия днРНК и микроРНК (Jin et al., 2020) (рисунок 1.3): (А) днРНК деградирует при связывании с микроРНК; (Б) днРНК выступает в качестве мишени (губки) для микроРНК; (В) днРНК конкурирует с микроРНК для регуляции уровня экспрессии мРНК; (Г) днРНК выступает предшественником микроРНК (сплайсируется в микроРНК). Более подробно разберем данные механизмы взаимодействия.

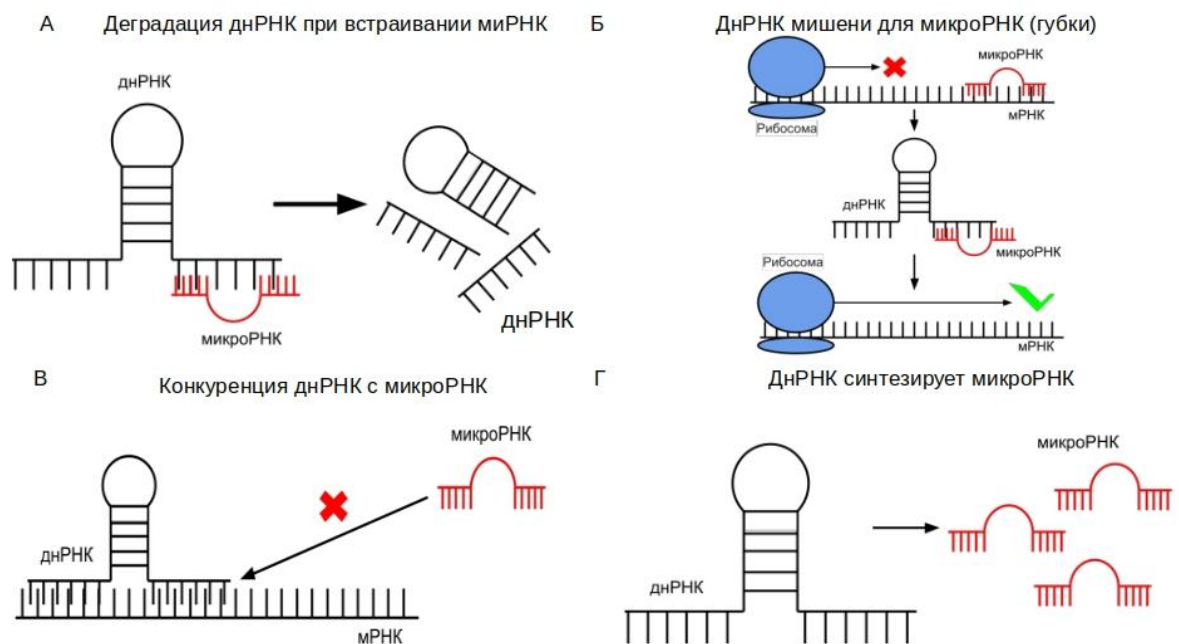


Рисунок 1.3. Механизмы взаимодействия днРНК-микроРНК: А) Дегградация днРНК при встраивании микроРНК. Б) днРНК выступает в качестве мишени (губки) для микроРНК. В) днРНК конкурирует с микроРНК для регуляции уровня экспрессии мРНК. Г) днРНК выступает предшественником микроРНК (сплайсируется в микроРНК).

Дегградация днРНК при встраивании микроРНК

МикроРНК функционируют в комплексе с белками семейства аргонавтов (AGO). Комплекс микроРНК-AGO связывается с комплементарными участками мРНК, что приводит либо к блокированию трансляции, либо к дегградации мРНК (Huntzinger и Izaurralde, 2011). Аналогичным образом мишенью для комплекса микроРНК могут выступать и последовательности днРНК. При связывании микроРНК с комплементарным участком днРНК происходит дегградация последней (Mukherjee et al., 2011; Yoon et al., 2013). В этом случае днРНК выступает как конкурент мРНК и тем самым адсорбирует часть молекул микроРНК, снижая их воздействие на молекулы потенциальных мишеней (мРНК). Важно, что при этом происходит дегградация и самой днРНК (рисунок 1.3А).

днРНК мишени для микроРНК

Впервые днРНК мишени с эффектом мимикрии, или конкурирующие эндогенные днРНК (competitive endogenous RNAs, ceRNAs), были обнаружены в геномах растений, рисунок 1.3Б (Thomson и Dinger, 2016). Механизм мимикрии («эффект губки») заключается в блокировании взаимодействия микроРНК с мРНК-мишенью за счет связывания данной микроРНК с днРНК-приманкой через частично комплементарные последовательности (Franco-Zorrilla et al., 2007).

На сегодняшний день экспериментально подтверждено влияние эндогенных днРНК на регуляцию синтеза фосфатов у *Arabidopsis thaliana* (Franco-Zorrilla et al., 2007), *Zea mays* (Du et al., 2018) и *Medicago truncatula* (Wang et al., 2017) за счет связывания с микроРНК. Методами биоинформатического анализа обнаружены эндогенные днРНК, принимающие участие в различных клеточных и молекулярных процессах (Unver и Tombuloglu, 2020), в управлении процессами роста и дифференцировки в организме, а также в развитии растения (Wang et al., 2015).

Конкуренция днРНК и микроРНК

Механизмы взаимодействия микроРНК с конкурирующими эндогенными днРНК схожи. Однако, если при «эффекте губки» днРНК выступает прямой

мишенью для микроРНК, то при конкурентном связывании (рисунок 1.3В) днРНК взаимодействует непосредственно с мРНК-мишенью. За счет комплементарного связывания с мРНК, днРНК не позволяет микроРНК связаться с мРНК-мишенью. Таким образом, в этом механизме днРНК конкурирует с микроРНК за сайт связывания на мРНК. Подобный механизм позволяет предотвратить деградацию мРНК, с которой связывается днРНК (Faghihi et al., 2010; Kimura et al., 2013; Wang et al., 2014).

днРНК предшественник микроРНК

Некоторые днРНК как первичные транскрипты могут служить предшественниками специфических микроРНК (рисунок 1.3Г). Малые РНК могут производиться за счет механизма сплайсинга и приводить к усилению посттранскрипционной регуляции соответствующих мРНК-мишеней (Ponting, Oliver и Reik, 2009). Такой механизм функционирования днРНК был экспериментально подтвержден у человека (Cai и Cullen, 2007). Показано, что они отвечают за регуляцию развития и прогрессирования опухоли за счет контроля транскрипции микроРНК-супрессоров (Augoff et al., 2012; Kallen et al., 2013). Для растений днРНК-предшественники обнаружены биоинформатическими методами. В работах Hirsch et al. (2006) и Amor et al. (2009) большинство выявленных микроРНК локализуются на концевых участках последовательностей днРНК, что указывает на то, что эти днРНК являются их предшественниками.

1.3.2 днРНК в альтернативном сплайсинге

За последние десятилетия стало ясно, что сплайсинг первичного транскрипта (пре-мРНК) является широко распространенным явлением, и что один ген может генерировать множество транскриптов, часто кодирующих разные белки, посредством процесса, известного как альтернативный сплайсинг (АС) (Rosenfeld et al., 1981). У эукариот аномалии альтернативного сплайсинга нередко приводят к болезням. У растений АС изучен хуже, однако показано, что он играет ключевую роль в регуляции протеома клетки растения в различных условиях, в том числе в стрессовых (Zorin et al., 2019).

днРНК способны действовать как регуляторные молекулы, модулирующие сплайсинг (рисунок 1.4), как косвенно, регулируя активность факторов сплайсинга, так и напрямую, взаимодействуя с пре-мРНК.

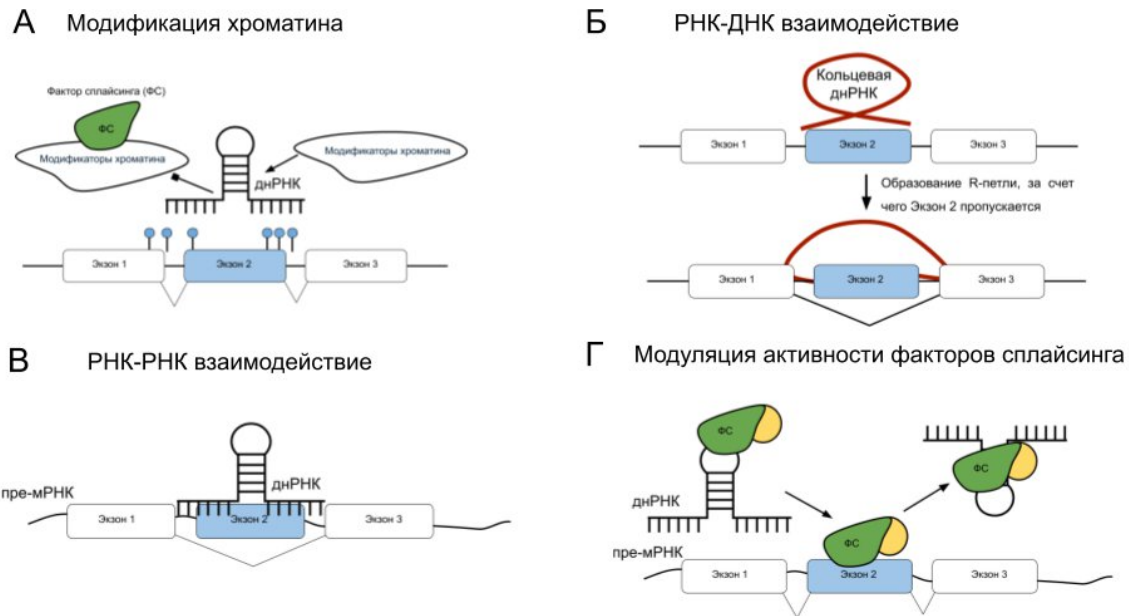


Рисунок 1.4. Влияние днРНК на альтернативный сплайсинг мРНК. А) днРНК регулируют модификации хроматина, рекрутируя или напрямую взаимодействуя с комплексами, модифицирующими гистоны, или ферментами в специфических хромосомных локусах. Б) Влияние на транскрипцию через образование R-петли при связывании кольцевой днРНК с локусом геномной ДНК. В) Образование РНК-РНК-дуплексов за счет гибридизации днРНК с пре-мРНК. Г) днРНК могут непосредственно воздействовать на активность и доступность факторов сплайсинга.

Регуляция альтернативного сплайсинга через модификацию хроматина

Модификации гистонов определяют доступность хроматина, который в конечном счете отражает активность транскрипции генов. Паттерны модификаций гистонов могут влиять на АС через хроматин-связывающий белок (связывает сайты сплайсинга в ДНК), расположенный внутри альтернативно сплайсируемого экзона, создавая препятствие для продолжения транскрипции, что способствует включению экзона в зрелую мРНК. В работах (Wang и Chang, 2011; Flynn и Chang, 2012) была показана связь днРНК с конформациями

хроматина. днРНК регулируют модификации хроматина, рекрутируя или напрямую взаимодействуя с комплексами, модифицирующими гистоны, или ферментами в специфических хромосомных локусах, которые, в свою очередь, модулируют транскрипцию генов (Tsai et al., 2010; Lucio et al., 2010).

Помимо модификации гистонов, трехмерная структура хроматина сама по себе играет важную роль в модулировании активности транскрипции при АС (Kornblihtt, 2006). В работе (Shukla et al., 2011) было показано, что СССТС-связывающий фактор (CTCF), являющийся трехмерным организатором хроматина, связывает участки ДНК-мишени, расположенные внутри альтернативно сплайсирующегося экзона, создавая препятствие для элонгации транскрипции, что способствует включению экзона в зрелую мРНК. днРНК взаимодействуют с фактором транскрипции CTCF через специфические сайты связывания (Amaral et al., 2018; Pisignano, 2019), способствуя образованию хромосомных петель. днРНК могут либо стабилизировать образованные петли, способствуя взаимодействию между энхансером и промотором конкретного экзона (что усиливает его транскрипцию), либо разрушать петли через вытеснение CTCF, блокируя доступ к регуляторным элементам.

Регуляция сплайсинга пре-мРНК через РНК-ДНК взаимодействие

Исследования показали потенциальную роль R-петли в альтернативном сплайсинге мРНК. R-петли представляют собой особые структурные конформации нуклеиновых кислот, которые образуются при гибридизации РНК с комплементарной цепью двухцепочечной ДНК. В результате вытесняется одна из исходных цепей ДНК в области гибридизации по типу петли (Drolet, 1994).

Длинные некодирующие РНК, а именно подкласс кольцевые РНК, образуются либо после выбора неправильного акцепторного сайта сплайсинга, либо в результате удаления нескольких последовательных экзонов в транскрипте при альтернативном сплайсинге (Nazirova, 2021). Таким образом, биогенез кольцевых РНК сам по себе может влиять на сплайсинг пре-мРНК и приводить к изменённым результатам экспрессии генов. В работе (Pisignano, 2021) показано, что кольцевые РНК также способны регулировать альтернативный сплайсинг через образование R-петли. После созревания кольцевая РНК способна

связываться с комплементарной цепью ДНК в своем родительском локусе, формируя стабильный РНК-ДНК-гибрид (R-петли). Возникновение R-петли создаёт препятствие для продвижения РНК-полимеразы II. Это замедляет или останавливает элонгацию, изменяя временное окно для сплайсинга. Например, кольцевая РНК, обнаруженная в *A. thaliana*, образует R-петлю при связывании с локусом геномной ДНК, что приводит к остановке транскрипции и способствует биогенезу вариантов гена за счёт альтернативного сплайсинга (механизм пропуска экзона) (Conn et al., 2017).

Регуляция сплайсинга пре-мРНК через РНК-РНК взаимодействие

днРНК могут влиять на регуляцию сплайсинга пре-мРНК через РНК-РНК взаимодействие. Независимо от своего геномного происхождения, антисмысловые днРНК могут гибридизировать с пре-мРНК и образовывать РНК-РНК дуплексы. Эти дуплексы способны маскировать сайты сплайсинга, препятствовать сборке сплайсосомы или, наоборот, привлекать факторы сплайсинга, что в итоге изменяет набор сплайсинговых изоформ (Pisignano, 2021).

Регуляция через модуляцию активности факторов сплайсинга

Помимо регуляции альтернативного сплайсинга через модификацию хроматина, днРНК могут непосредственно воздействовать на активность и доступность факторов сплайсинга (ФС) (Pisignano и Ladomery, 2021). Данный механизм основан на формировании комплексов ФС-днРНК. Эти комплексы влияют на посттрансляционные модификации ФС, их внутриядерную локализацию и, как следствие, на эффективность взаимодействия ФС с пре-мРНК (Hutchinson et al., 2007).

В механизме регуляции альтернативного сплайсинга через модуляцию ФС принимают участие днРНК, имеющие различные молекулярные функции. Важно отметить, что днРНК в данном механизме способны выполнять сразу несколько молекулярных функций. Например, днРНК MALAT1 может выступать в качестве направляющей для ФС, образуя с ним комплекс и доставляя к месту сплайсинга пре-мРНК (Bernard et al., 2010). Также данная днРНК способна выступать в качестве каркаса для сборки мультибелковых комплексов, способствуя

рекрутированию ФС к определённым участкам пре-мРНК, тем самым усиливая или подавляя использование тех или иных сайтов сплайсинга (Malakar et al., 2017).

Ещё одним механизмом модуляции является конкуренция днРНК и пре-мРНК за связывание с ФС (Porto, 2019). днРНК содержит последовательности, комплементарные сайтам сплайсинга на пре-мРНК, что приводит к их маскировке и блокировке связывания факторов сплайсинга. Это приводит к блокированию сайтов связывания ФС и изменению выбора альтернативных экзонов или сайтов сплайсинга (Bardou et al., 2014; Porto, 2019).

1.4 Экспрессия днРНК

На сегодняшний день появляется все больше работ по анализу днРНК у различных организмов, в которых применяются методы исследования экспрессии данных молекул. Общепринятым считается, что днРНК демонстрируют более низкий уровень экспрессии по сравнению с мРНК, при этом экспрессируются более специфично для определенных тканей (Derrien et al., 2012). Однако в последнее время начинают появляться работы, демонстрирующие днРНК с уровнями экспрессии, равными таковым для белок кодирующих генов (De Quattro, 2017; Lamin-Samu et al., 2022). Подобные случаи являются единичными, и их результаты требуют экспериментального подтверждения.

Выделяют три направления исследования экспрессии днРНК: профилирование базового уровня экспрессии; анализ дифференциальной экспрессии и коэкспрессии; анализ тканеспецифичности. Данные подходы применяются как для масштабного транскриптомного анализа, так и для анализа небольших выборок последовательностей.

Профилирование базового уровня экспрессии днРНК является наиболее распространенным направлением исследования. Оно заключается в количественной оценке транскрипционной активности днРНК в различных тканях, на определённых стадиях растений или в ответ на внешние стимулы. Основным результатом данного этапа служит матрица экспрессии, содержащая как абсолютные, так и нормированные значения. Большинство работ начинают

свое исследование с данного направления, постепенно углубляясь в дифференциальную экспрессию или коэкспрессию.

Анализ дифференциальной экспрессии и коэкспрессии — статистическое сравнение уровней экспрессии между контрастными группами образцов для выявления достоверно изменяющихся транскриптов. Параллельно оценивается корреляция динамики днРНК с белок кодирующими генами, что позволяет судить об их функциональных взаимодействиях (Yang, 2019).

Анализ тканеспецифичности днРНК является наименее исследованным, так как нет точного понимания, какой метод определения специфичности экспрессии гена в тканях наиболее точен (Kryuchkova-Mostacci, 2017). Во многих работах тканеспецифичным считается ген, имеющий экспрессию только в одной ткани (Li et al., 2014; De Quattro, 2017; Han et al., 2019). Одним из первых подходов является подсчет количества тканей, в которых данный ген экспрессируется (Subramanian и Kumar, 2004). Однако этот метод позволяет исследовать только гены с высокой экспрессией, что затрудняет выбор порога фильтрации последовательностей. Также существует проблема близкородственных тканей, таких как части стебля или мозга человека.

С развитием понимания о структуре и функциях днРНК стало появляться все больше подходов к решению данной задачи. Существующие методы делятся на две группы. В первой группе показателем тканеспецифичности является число, определяющее, насколько неравномерно распределена экспрессия гена по всем исследованным тканям: 1 - если ген имеет специфическую экспрессию, 0 - если экспрессируется в нескольких тканях - Tau (Yanai et al., 2005), Gini (Ceriani и Verme, 2012), tissue specificity index (TSI) (Julien et al., 2012). Во второй группе показателем тканеспецифичности является отношение уровня экспрессии гена в определенной ткани относительно его уровня экспрессии в других тканях. К метрикам данной группы относятся - z-score, SPM (Xiao et al., 2010), expression enrichment (EE) (Yu et al., 2006).

Важно отметить, что, хотя обе группы методов оценивают тканеспецифичность, они отражают разные аспекты этого свойства. Метрики первой группы носят больше глобальный характер, они характеризуют общее

распределение экспрессии гена по всем исследуемым тканям, сводя его к единому показателю. В отличие от них, метрики второй группы ориентированы на детальный анализ и рассчитывают степень специфичности гена индивидуально для каждой конкретной ткани.

Однако несмотря на различия в подходах, в основном все метрики расчета тканеспецифичности гена (Tau, EE score, TSI, Gini coefficient) имеют похожие входные данные. Для расчета используется уровень экспрессии гена в определенной ткани и количество экспериментов, которым данная ткань принадлежит. Например, формула для расчета Tau (Yanai et al., 2005):

$$\tau = \frac{\sum_{i=1}^n (1 - \hat{x}_i)}{n-1}, \quad (1)$$

$$\hat{x}_i = \frac{x_i}{\max_{1 \leq i \leq n} (x_i)}, \quad (2)$$

где, x_i - экспрессия гена в ткани, n - количество экспериментов с тканью.

Помимо перечисленных выше метрик встречаются работы, использующие метрики, основанные на энтропии — Шеннона, Jensen-shannon (JS) (Huang, 2005):

$$H_g = \sum_{1 \leq t \leq N} -p_{t|g} \log_2(p_{t|g}), \quad (3)$$

$$p_{t|g} = \frac{\omega_{g,t}}{\sum_{1 \leq t \leq N} \omega_{g,t}}, \quad (4)$$

где $p_{t|g}$ — распределение экспрессии гена в ткани и $\omega_{g,t}$ - экспрессия гена. Таким образом $p_{t|g}$ показывает отношение экспрессии гена в определенной ткани на сумму экспрессий этого гена во всех тканях.

$$Q_{g|t} = H_g - \log_2(p_{t|g}), \quad (5)$$

где H_g — энтропия, показывает, насколько ген тканеспецифичный. Если значение энтропии близко или равно 0, то данный ген специфичен определенной ткани. $Q_{g|t}$ - показывает энтропию экспрессии исследуемого гена в одной ткани.

1.5 Биологическая роль днРНК у растений

Предыдущие разделы показали разнообразие и важность молекулярных функций днРНК у растений. днРНК могут выступать как генетические и эпигенетические регуляторы экспрессии генов. В зависимости от пространственной локализации действия днРНК могут функционировать как цис-регуляторные молекулы, взаимодействуя с близко расположенными генами в том же локусе, или как транс-регуляторные молекулы, действуя на гены в удаленных локусах или других хромосомах. днРНК могут препятствовать связыванию транскрипционных факторов с промоторной областью, выступая в качестве репрессоров транскрипции. В результате этих молекулярных процессов контролируются функции растений на более высоком уровне организации: тканей, органов и всего организма. Примеры процессов, которые затрагивает регуляция при помощи днРНК у растений, приведены на рисунке 1.5. К ним относятся: устойчивость растений к различным видам стресса, развитие вегетативных органов растений, яровизация, развитие плода.

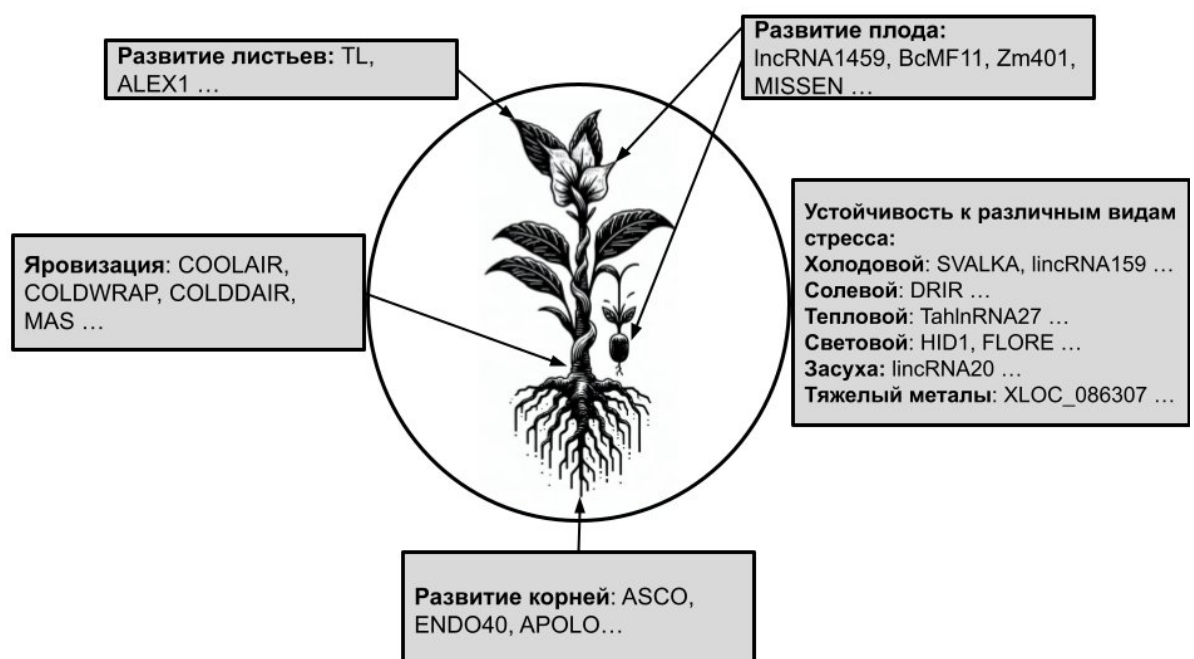


Рисунок 1.5. Биологическая роль днРНК у растений. днРНК принимают участие в регуляции генов, отвечающих за устойчивость растений к различным видам стресса. Также они влияют на плодовитость растений, принимают участие в развитии листа, корня и в процессах яровизации.

днРНК в процессе яровизации

Цветение является одним из важнейших адаптивных признаков, обеспечивающих переход репродуктивного роста и развития, происходящий в благоприятных условиях в течение жизненного цикла растения. Яровизация - переход растений из вегетативного состояния в генеративное, связанное с формированием зачатков цветков в конусах нарастания. Этот переход индуцируется холодом. Яровизация играет важную роль в образовании цветков у озимых (прорастающих осенью и зимующих в стадии проростков) и у многолетних растений (Marquardt et al., 2014). Яровизация также является наиболее изученным регуляторным процессом у растений, в котором участвуют днРНК. днРНК принимают участие в регуляции гена *FLOWERING LOCUS C (FLC)*, отвечающего за устойчивость *A. thaliana* к холодовому стрессу (Zhu и Wang, 2012). На сегодняшний день в локусе *FLC* обнаружено два класса днРНК: *COOLAIR* - днРНК, транскрибируемая в антисмысловом направлении по отношению к промотору гена *FLC*; *COLDAIR* - днРНК, транскрибируемая с первого интрона гена *FLC* в смысловом направлении. днРНК *COOLAIR* связываются с локусом *FLC* в цис-положении и образуют R-петлю между промотором и первым интроном гена *FLC*, что приводит к подавлению экспрессии данного гена-мишени.

днРНК и влияние на урожайность растений

Развитие нормального пыльцевого зерна имеет решающее значение для оплодотворения и образования семян. Пыльник действует как родительская ткань пыльцевого зерна и регулирует его развитие. Аномальное пыльцевое зерно может привести к развитию растений с мужской стерильностью (Liu et al., 2015). Экспериментальные исследования (Ding et al., 2012; Song, Cao и Wang, 2013) показали вовлеченность днРНК в процесс контроля урожайности. Например,

днРНК *LDMAR* риса (*Oryza sativa japonica*) вызывает запрограммированную гибель клеток пыльников в условиях длинного дня (*LD*, long day) за счет сплайсинга с образованием микроРНК *osa-smR5864*. Однонуклеотидная мутация в микроРНК *osa-smR5864* приводит к мужской стерильности растения (Ding et al., 2012). Снижение экспрессии днРНК *BcMF11* у *Brassica campestris* препятствует деградации тапетума и приводит к нарушению образования пыльцевых зерен (Song, 2013). Также методами биоинформатики в работе (Huang et al., 2018) для пяти стадий развития пыльника *Brassica rapa* было выявлено 14 днРНК, коэкспрессирующихся с 10 известными генами, кодирующими белки, которые играют решающую роль в развитии пыльцы. Как утверждают авторы, выявленные днРНК представляют интерес для дальнейших экспериментальных исследований вовлеченности днРНК в репродуктивные процессы растений.

днРНК в процессе фотоморфогенеза

Молекулярный сигнальный механизм фотоморфогенеза широко изучался у различных видов растений. В этом процессе ряд белков выступает в качестве первичных и вторичных сигнальных молекул. Среди них выделяют семейство транскрипционных факторов со структурой ДНК-связывающего домена «спираль-петля-спираль» (helix-loop-helix, bHLH). Среди членов семейства выделяют факторы взаимодействия с фитохромом (PIF), которые выполняют функцию подавления фотоморфогенеза проростков в темное время суток (Kim et al., 2003). В работе (Wang et al., 2014) обнаружена днРНК *HID1*, которая функционирует как транс-регуляторный элемент, взаимодействуя с промотором гена *PIF3* (ключевая сигнальная молекула фотоморфогенеза, негативно регулирующая реакции на красный свет) и ингибирует его транскрипцию. В результате наблюдается удлинение зародышевого стебля (гипокотилия) у проростков *A. thaliana*.

днРНК в процессах устойчивости к различным видам стресса

В природе растения подвергаются действию различных стрессовых факторов, таких как соль, засуха, холод, тепло, а также различные вирусные инфекции. Это приводит к проблемам в росте и урожайности растений. Чтобы адаптироваться и выжить в таких неблагоприятных условиях, растения

используют множество механизмов регуляции генов для восстановления гомеостаза. Современные исследования демонстрируют немаловажную роль днРНК в регуляции экспрессии генов в ответ на стрессовые состояния (Wang et al., 2017). Вовлеченность днРНК в процессы устойчивости к абиотическим видам стресса показана у *A. thaliana* (Amor et al., 2009), *B. rapa* (Wang et al., 2019), винограда (*Cabernet Sauvignon*) (Wang et al., 2019), *O. sativa* (Chung et al., 2016) и многих других растений. Например, у *A. thaliana* в регуляции устойчивости к холодовому стрессу участвует несколько механизмов действия днРНК, функционирующих преимущественно как цис-регуляторные элементы:

- Модуляция хроматина. Например, днРНК *COLDAIR* (транскрипция которой была описана выше), привлекает к этому участку хроматина белки-репрессоры группы Polycomb (PRC2). Те, в свою очередь, ставят на гистонах метку H3K27me3, и транскрипция *FLC* замолкает.
- Альтернативный сплайсинг. днРНК *TAS1* приводит к образованию варианта транскрипта с нерасщепленным интроном, более устойчивого к холодовому стрессу, чем сплайсированный вариант данной днРНК (Calixto et al., 2019).
- Транскрипционная интерференция. Например, днРНК *SVALKa* подавляет транскрипцию гена *CBF1*, ответственного за запуск каскада реакций холодоустойчивости. Это происходит за счёт того, что гены *SVALKa* и *CBF1* транскрибируются одновременно с одного и того же участка ДНК, но в противоположных направлениях. В результате их РНК-полимеразы движутся навстречу друг другу, происходит транскрипционная коллизия, которая обрывает синтез мРНК *CBF1* (Jha et al., 2020).

днРНК *lncRNA973* у *A. thaliana* регулирует устойчивость растений к солевому стрессу путем модуляции экспрессии генов, ответственных за активные формы кислорода (*SOD*, *CAT* и *POD*), транскрипционные факторы (*MYB5*, *WRKY46*, *NAC29* и *ERF62*) и ионный баланс (*NHX7*) (Zhang et al., 2019). Сверхэкспрессия днРНК *At5NC056820* у *A. thaliana* повышает устойчивость

растения к засухе (Wang et al., 2023). днРНК *TCONS_00021861* у *O. sativa* за счет эффекта губки взаимодействует с *miR528-3p* и предотвращает её взаимодействие с геном-мишенью *YUCCA-7* (который участвует в биосинтезе триптофан-зависимых ауксинов и повышает устойчивость к засухе), что активирует синтез индолил-3-уксусной кислоты (ИУК), повышающей устойчивость к засухе (Chen, 2021). Подобный механизм наблюдается для днРНК *TCONS_00048391*, которая является «губкой» для микроРНК *bra-miR164a*, подавляющей экспрессию гена-мишени *NAC1* (ответственного за устойчивость к тепловому стрессу у *B. rapa*) (Wang et al., 2019). Таким образом, даже внутри одного организма днРНК имеют различные механизмы регуляции устойчивости к абиотическим видам стресса.

1.6 Эволюция днРНК

Эволюционные процессы приводят к изменчивости структуры и функций днРНК. В области изучения эволюции днРНК существует несколько направлений активных исследований. Прежде всего это вопросы, связанные с возникновением новых днРНК в геномах организмов. Одним из таких механизмов является дупликация генов. Появление в геноме дополнительной копии гена открывает свободу эволюции одной из копий. Мутации, возникающие в одной из двух копий и ослабляющие исходную функцию гена, не будут отсеиваться отбором, потому что вторая копия может сохранить прежнюю функциональность. Таким образом, за счет дупликации возникают «псевдогены» (неактивные участки), изоформы и транскрипты днРНК (Khorikova, 2015).

Существует несколько гипотез возникновения днРНК за счет дупликации. Первая основывается на потере одной из копий белок кодирующих генов (в процессе дупликации), способности кодировать белок. Это объясняется наличием у днРНК регуляторных элементов, сайтов сплайсинга, последовательностей экзонов и последовательностей, имеющих поли(А)-хвост. Наиболее известный пример днРНК, возникшей за счет дупликации гена, кодирующего белок, — человеческая днРНК *Xist* (Duret et al., 2006). Вторая гипотеза основана на дупликации некодирующих РНК. Однако подобные случаи, в которых днРНК возникают из дублированных копий некодирующих РНК, встречаются

значительно реже, чем случаи их происхождения от дублированных белок кодирующих генов. Это объясняется высокой эволюционной лабильностью днРНК (Kazemzadeh, 2015).

Также одна из гипотез возникновения днРНК связана со встраиванием мобильных элементов (МЭ) в последовательность транскрипта. Мобильные элементы могут способствовать регуляции экспрессии генов при развитии и адаптации, а также могут служить основными источниками генетических вариаций в эволюции геномов. МЭ присутствуют в экзонах, сайтах начала транскрипции, сайтах полиаденилирования (поли(А)-хвосты) или в некоторых комбинациях этих частей. Исследование, проведенное на 8 растениях (включающих *A. thaliana*, кукурузу, томат, тополь) (Zhu et al., 2022b), показало, что у 59.7 % днРНК, найденных во всех восьми организмах, как минимум один экзон перекрывается с последовательностью МЭ. В исследовании (Kapusta et al., 2013) показано участие МЭ не только в появлении днРНК, но и в регуляции и вариативности днРНК.

Для оценки консервативности днРНК между геномами различных организмов сравнивают их последовательности (Ulitsky, 2016). Например, при выравнивании 67 тысяч днРНК риса на 39 тысяч днРНК кукурузы доля совпадающих позиций гомологов днРНК равна 20 % (Wang et al., 2015). При сравнении днРНК для эволюционно более далеких видов, например, *A. thaliana* и *O. sativa*, процент идентичности значительно падает вплоть до отсутствия гомологии (Zhu et al., 2022). Однако некоторые последовательности днРНК демонстрирует уровень консервативности, сравнимый с консервативностью белок кодирующих генов (Nitsche и Stadler, 2017). Одной из таких днРНК является межгенная *IPS*, участвующая в гомеостазе фосфора (P_i) у *A. thaliana*, *O. sativa*, *S. lycopersicum* и *S. moellendorffii*. *IPS* содержит консервативный мотив из 24 нуклеотидов у перечисленных видов, который служит мишенью для *miR399*. *IPS* первоначально появилась у *S. moellendorffii*, представителя древних споровых высших растений. За время эволюционного развития функция *IPS* не изменилась, что может указывать на роль данной днРНК в адаптации к суше в процессе миграции из водной среды.

Однако большинство днРНК показывает высокую вариабельность первичной структуры, что подтверждается в работе (Sang et al., 2021), посвященной анализу консервативности днРНК между 25 цветковыми растениями. Анализ показал, что только 6.79 % днРНК консервативны у всех исследуемых растений. В работе (Zhang et al., 2014) был проведен анализ консервативности днРНК (межгенных) риса относительно семи видов растений. Авторы показали, что для межгенных днРНК риса доля выровненных нуклеотидов относительно ортологичных участков других видов, составила 18 % для интронов и 23 % для экзонов и оказалась близкой с таковой у межгенных участков (20 %), выше, чем у мРНК мобильных элементов (12 % для интронов и 10 % для экзонов), но гораздо меньше, чем у мРНК (56 % в интронах и 86 % в экзонах). Анализ консервативности на основе оценки доли нуклеотидных замен также показал низкую консервативность днРНК по сравнению с мРНК, однако более высокую по сравнению с мРНК мобильных элементов. Анализ консервативности структурных элементов днРНК показал высокую консервативность экзонов днРНК по сравнению с интронами. В данных исследованиях показана быстрая эволюция днРНК, которая приводит к низкой консервативности последовательностей. Также отмечается, что процент консервативных днРНК выше среди родственных видов (в пределах одного рода), при этом консервативность резко убывает по мере увеличения эволюционного расстояния (Sang et al., 2021).

Несмотря на существующие теории возникновения и эволюции днРНК, строгих доказательств для них пока не существует. Однако в целом анализ показывает, что они более вариабельны по сравнению с мРНК.

1.7 Методы секвенирования для исследования днРНК

До недавнего времени идентификация днРНК считалась побочным результатом исследований, направленных на выявление белок кодирующих генов. Однако за последнее десятилетие стало очевидно, насколько важна роль днРНК в молекулярных и биологических процессах различных организмов. Это привело к увеличению количества экспериментов, направленных на

секвенирование и дальнейшую аннотацию днРНК. На сегодняшний день методы секвенирования днРНК можно разделить на две группы:

1. RNA-seq - наиболее распространённый метод секвенирования днРНК. В зависимости от способа удаления нецелевых фракций различают два основных подхода поли(А)-селекция и деплеция рРНК.
2. GRO-seq - методы, позволяющие секвенировать РНК во время её синтеза РНК-полимеразой, независимо от того, будет ли эта молекула разрушена в клетке впоследствии или нет. К данным методам относится также NET-seq, plaNET-seq.

Использование РНК-секвенирования (RNA-seq) является наиболее часто используемым методом для массовой идентификации транскриптов благодаря простоте и возможности выявления новых молекул без предварительного знания их последовательности (Cheung et al., 2006). Однако при секвенировании днРНК этот метод сталкивается с рядом ограничений. днРНК часто имеют несколько копий транскриптов, которые могут функционировать по-разному в зависимости от их положения относительно генов-мишеней (Ponting и Reik, 2009). Поэтому учёт днРНК на уровне отдельных транскриптов имеет важное значение. Одним из недостатков метода RNA-seq является неравномерное покрытие транскриптов прочтениями, особенно для редких изоформ. В результате исследователи получают неполный набор транскриптов организма (Deveson et al., 2017). Кроме того, наиболее часто используемые протоколы RNA-seq направлены прежде всего на идентификацию молекул мРНК, содержащих поли(А)-хвост. Поэтому они используют селекцию РНК по этой последовательности для обогащения полиаденилированных молекул, что приводит к потере неполиаденилированных днРНК. Однако во многих работах показано, что более 70% днРНК растений транскрибируется РНК-полимеразой II и полиаденилирована (Mattick et al., 2023; Ziegler, 2017; Kornienko et al., 2024). Таким образом, поли(А)-селекция не исключает днРНК из анализа полностью. Именно поэтому многие крупномасштабные исследования, в том числе и настоящая работа, используют поли(А)-селекцию библиотек, принимая этот компромисс ради возможности

анализировать сотни индивидуальных генотипов и получать широкую картину генетического разнообразия днРНК.

Для более полного охвата некодирующих транскриптов в рамках RNA-seq применяют также стратегию деплеции рРНК. В отличие от протоколов RNA-seq, в которых для секвенирования мРНК используют селекцию по поли(А)-хвосту, метод деплеции рРНК основан на гибридизации специальных олигонуклеотидных зондов, комплементарных различным видам рРНК (Dahlgren, 2020). Эти зонды связываются с соответствующими молекулами рРНК, образуя двухцепочечные комплексы, которые затем удаляют из образца. В результате из набора данных удаляются все рРНК без вреда для остальных некодирующих РНК. Данный метод чаще используют для получения наиболее полного набора днРНК в конкретном образце, однако он не всегда демонстрирует значимый прирост числа секвенированных днРНК по сравнению с поли(А)-селекцией и проигрывает ей в технологической простоте при массовом секвенировании. Например, исследование, проведённое (Dahlgren et al., 2020) и направленное на сравнение поли(А)-селекции и деплеции рРНК на образцах ткани печени и мозга лошадей, показало, что поли(А)-селекция захватывает больше днРНК в обеих тканях по сравнению с деплецией рРНК. Авторы предполагают, что некоторые неполиаденилированные днРНК могут стабилизироваться вторичными структурами, препятствующими их изоляции во время процедуры обеднения рибосомальной РНК.

Для преодоления ограничений RNA-seq были разработаны методы, основанные на секвенировании РНК во время её синтеза РНК-полимеразой. Метод GRO-seq позволяет выявлять РНК в процессе синтеза, включая транскрипты, подвергающиеся быстрой деградации (Lopes, 2017). В исследованиях, проведённых с использованием GRO-seq, отмечают высокую чувствительность при выявлении низкоэкспрессирующихся транскриптов и возможность обнаружить все синтезирующиеся РНК одинаково, независимо от их стабильности в клетке (Lopes, 2017; Zhu et al., 2018; Hetzel et al., 2016). Это делает метод особенно подходящим для изучения некодирующих РНК с низкой экспрессией и высокой скоростью деградации. Методы NET-seq и plaNET-seq

основаны на селекции РНК-полимераз через FLAG-метки (Churchman, 2012). Преимущество данного подхода заключается в том, что при захвате полимеразы одновременно извлекается и синтезируемая ею РНК (Huffines, 2025). Таким образом, за счёт этих методов появляется возможность выявить все некодирующие транскрипты в момент их синтеза, включая нестабильные, короткие и неполиаденилированные молекулы, которые пропускаются стандартным RNA-seq.

Несмотря на перечисленные преимущества описанных методов в секвенировании днРНК, RNA-seq по-прежнему остаётся наиболее используемым методом в работах, связанных с исследованием данных молекул (Mattick et al., 2023; Ziegler, 2017; Kornienko et al., 2025). Это связано с трудоёмкостью экспериментов, проводимых с помощью технологий GRO-seq, NET-seq и подобных. Данные методы требуют аккуратного выделения и обогащения ядер и тканей, что повышает риск возникновения артефактов при секвенировании (Wang, 2019). Также отмечается (Hetzl et al., 2016), что данные методы хорошо разработаны для человека и млекопитающих, но для растений редко применяются вследствие технических сложностей изоляции ядер из растительных клеток и высокого содержания хлоропластов, которые загрязняют образец. В связи с этим они не применяются для массового секвенирования и всё ещё остаются редкими методами в исследовании растений (Wang, 2019; Hetzel et al., 2016).

На сегодняшний день существует достаточно большое количество методов секвенирования днРНК. Однако они отличаются областью применения. Если целью исследования является получение наиболее полного набора днРНК в определённой ткани одного организма, рекомендуется использовать методы GRO-seq. Если же целью работы является исследование большого количества организмов, чаще используют методы RNA-seq, принимая во внимание возможную потерю информации.

1.8 Биоинформатические методы идентификации днРНК

Основной задачей в идентификации днРНК является точное разделение выборки РНК на кодирующие белок и некодирующие. В предыдущем разделе

было показано, что современные подходы к идентификации днРНК базируются на массовом секвенировании РНК, что приводит к появлению значительных объёмов данных. В связи с этим возникает необходимость в применении методов биоинформатики.

Современные биоинформатические подходы используют различные методы идентификации днРНК (Zhang и Zhou, 2014; De Quattro и Bertolini, 2017; Zhao et al., 2018), рисунок 1.6. К ним относятся как классические методы, основанные на анализе структурных и композиционных особенностей последовательностей, так и методы машинного и глубокого обучения. К программам, использующим классические подходы для идентификации днРНК, относятся CPC (Kong et al., 2007) и LncRNA-ID (Achawanantakun et al., 2015). Основной их задачей является вычисление кодирующего потенциала исследуемого транскрипта. Основными критериями, по которым происходит оценка кодирующего потенциала, являются признаки открытых рамок считывания (ОРС) - их длина, покрытие и целостность (Fickett и Tung, 1992); расчет параметра Fickett, вычисляемый на основе сравнения позиционных частот нуклеотидов в кодонах с эталонными распределениями для белок кодирующих генов (Fickett, 1982); характеристики GC-состава (Pozzoli et al., 2008), а также сравнение структурных особенностей транскриптов, таких как длина транскрипта и количество экзонов/интронов (Sun et al., 2013). Однако наличие общих характеристик днРНК и белок кодирующих генов может приводить к некорректному предсказанию. Например, встречаются случаи сходства признаков ОРС днРНК с ОРС белок кодирующих генов, однако покрытие или целостность ОРС обычно у днРНК хуже (Dinger et al., 2008).



Рисунок 1.6. Методы предсказания длинных некодирующих РНК.

Следующий блок программ использует методы машинного обучения для предсказания днРНК. К подобным программам относятся: lncFinder (Han et al., 2019), CPAT (Wang et al., 2013), CPC2 (Kang et al., 2017), PLEK (Zhang и Zhou, 2014), CNCI (Sun et al., 2013). Данные программы используют сразу несколько альтернативных наборов признаков для выявления днРНК. На сегодняшний день выделяют три блока признаков: структурные свойства, особенности вторичной структуры и физико-химические свойства РНК (Han et al., 2019). Использование подобных признаков позволяет различать РНК сразу на нескольких уровнях, что повышает качество предсказания. Помимо разнообразия извлекаемых признаков, перечисленные программы используют различные модели машинного обучения. Так, например, программа CPAT применяет метод логистической регрессии для построения модели оценки кодирующего потенциала. Программа PLEK использует метод опорных векторов для более точного подбора k-меров с целью дальнейшего разделения выборки на мРНК и днРНК. LncFinder использует сразу несколько методов машинного обучения (Random Forest, логистическая регрессия, метод опорных векторов, Gradient Boosting), из которых затем выбирает наилучшую модель для дальнейшего предсказания.

Методы глубокого обучения представляют собой современное направление в идентификации днРНК, способное автоматически выявлять

сложные паттерны и зависимости в данных без необходимости ручного подбора признаков. К программам, использующим данный подход, относятся: LncADeep (Yang et al., 2018), LncDLSM (Wang et al., 2023), lncRNAet (Baek et al., 2018). Преимуществом методов глубокого обучения над методами машинного обучения является наличие многослойной архитектуры нейронных сетей. Это позволяет выявлять нелинейные зависимости и взаимодействия между признаками. Например, программа lncRNAet использует архитектуру рекуррентной нейронной сети для моделирования последовательностей РНК, что позволяет выявлять внутренние характеристики транскриптов днРНК. Программа LncADeep использует архитектуру глубокой сети доверия (DBN), в которую интегрируются два типа признаков, основанные на структурных признаках РНК и их гомологии друг с другом. Подход, использующий большее количество признаков в комбинации с возможностью нейронной сети обучаться новым ассоциациям между признаками, дает значительное повышение качества предсказания по сравнению с предыдущими методами.

Однако само выявление днРНК из исследуемой выборки не несет полезной информации о функциях, механизмах действия или биологической значимости данных молекул в организме (Guo et al., 2013). Таким образом, функциональная аннотация является обязательным критерием в исследовании биологической значимости днРНК. Аннотация включает: классификацию днРНК на основе их положения относительно белок кодирующих генов, анализ структурных особенностей (длины транскрипта, количества экзонов/интронов), анализ экспрессии, коэкспрессии с другими днРНК и белок кодирующими генами, поиск гомологичных последовательностей с другими организмами, оценку видоспецифичности. Классификация днРНК разделяет выявленные некодирующие РНК на четыре группы: антисмысловые, смысловые, интронные и межгенные. Функции данных классов днРНК описаны в пункте 1.2 «Классификация днРНК по положению в геноме». Анализ структурных особенностей днРНК в основном направлен на оценку длины транскрипта, количества экзонов и интронов, а также хромосомной локализации. По видоспецифичности днРНК разделяют на две группы: поливидовыми считаются

днРНК с уровнем идентичности выше 50 % с днРНК других организмов, внутривидовыми - с отсутствием гомологов с другими организмами (Kang et al., 2017). Источником информации об этих последовательностях служат базы данных. Для их пополнения используются различные источники: результаты биоинформатической обработки транскриптомных экспериментов и предсказания днРНК, экспериментальные результаты, полученные при анализе литературных источников, совместное использование как биоинформатических методов, так и литературных источников. Базы данных, описывающие структуру и функции днРНК растений, приведены в таблице П.1.

Таким образом, на сегодняшний день существуют десятки инструментов и программ предсказания днРНК, основанных на технологиях машинного и глубокого обучения и использующих различные функции. Несмотря на то, что перечисленные вычислительные методы показали высокие результаты в идентификации днРНК, насколько нам известно, не существует единого мнения относительно эффективности этих инструментов и признаков, позволяющих однозначно определить днРНК из набора РНК.

1.9 Пан-геномы и пан-транскриптомы

1.9.1 Пан-геномы

В пределах одного биологического вида существует высокое геномное и морфологическое разнообразие, обусловленное структурными вариациями. Они включают в себя инсерции, дупликации, инверсии, транслокации, а также вариации числа копий генов и их наличия или отсутствия. Данные вариации способны влиять на сегменты ДНК длиной в несколько мегабаз и в некоторых случаях изменять число хромосом. Вследствие этого один референсный геном не способен охватить все генетическое разнообразие вида. Например, в референсных геномах растений отсутствует ряд генов, важных для сельскохозяйственных исследований. Для решения данной проблемы используется пан-геном.

Термин пан-геном был впервые сформулирован в работе, посвященной бактериальному виду *Streptococcus agalactiae*. На сегодняшний день существует несколько определений термина «пан-геном»:

1. Общее число неизбыточных генов, представленных в отдельно взятом биологическом виде (Golicz et al., 2016).
2. Полный набор генов одного вида (Meile et al., 2018).
3. Результаты геномного сравнения различных организмов одного вида или рода (Alcaraz et al., 2010).
4. Общий набор генов, представленный в группе близких видов или штаммов одного вида (Rasko et al., 2008).
5. Совокупность генов всех живых организмов, вирусов и различных мобильных генетических элементов (Pang et al., 2005).

В работе (Merot-L'anthoene et al., 2019) выделяют два понимания термина пан-генома: на функциональном основании и на структурном основании:

- Структурное понимание утверждает, что пан-геном является полным набором неизбыточных генов длиной примерно 100 п.н. (возможно, длиннее в зависимости от особи). Преимущество данного понимания заключается в принятии во внимание как последовательностей генов, так и негенных последовательностей.
- Функциональное понимание заключается в том, что пан-геном является совокупностью всех генов в данном наборе индивидуумов. Он может быть расширен на уровне семейства генов, при этом в данном понимании особо схожие гены могут приниматься за один. Таким образом, если хотя бы один представитель семейства генов присутствует в каждом представителе вида, то он относится к основному геному.

В данной работе под пан-геномом понимается полный неизбыточный набор генов, представленных в геноме хотя бы одной особи, относящейся к данной таксономической единице (Propozin, 2021). Отметим, что это понятие можно распространить на таксономические единицы любого порядка (Vernikos et al., 2015). Так, была предпринята попытка создать общий пан-геном бактерий на основе 573 секвенированных ранее геномов бактерий (Lapierre и Gogarten, 2009). Однако в настоящее время понятие пан-геном чаще всего употребляют именно в отношении вида.

Гены пан-генома могут быть разделены на две группы, рисунок 1.7:

1. Ядро - гены, которые присутствуют в геномах всех представителей таксономической единицы.
2. Оболочка и облако - гены, которые присутствуют в геномах не всех представителей вида. Они, в свою очередь, делятся на два типа: гены, представленные в 99-10% всех представителей вида (оболочка), и гены, представленные в <10% всех представителей вида (облако).

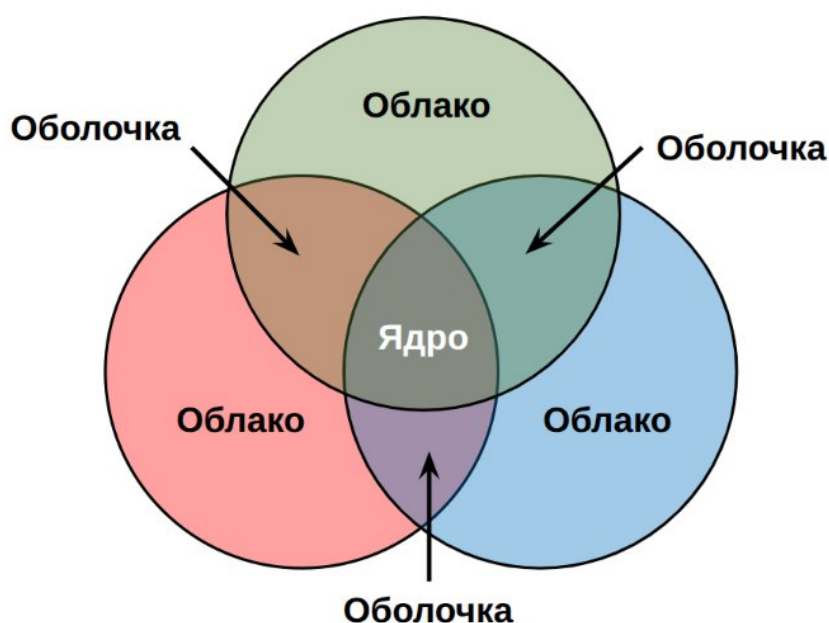


Рисунок 1.7. Типы генов в пан-геноме. Ядро, Оболочка, Облако.

При увеличении количества геномов, включенных в пан-геном, общее число генов в каждой из перечисленных групп в этом пан-геноме будет возрастать. Причем каждая из этих групп генов рано или поздно выйдет на плато, после которого она не будет значимо изменяться с добавлением новых геномов. Отмечается, что гены оболочки и облака являются переменными и в среднем эволюционно более молоды, чем гены ядра (Golicz et al., 2016). Этот результат кажется логичным в свете того, что большая часть генов ядра является консервативной и составляют гены домашнего хозяйства. Функции генов ядра и оболочки/облака различаются: функции ядра связаны с важнейшими клеточными процессами, а функции оболочки/облака - с окружающей средой и защитными реакциями, рецепторной и антиоксидантной активностью, регуляцией генов и

передачей сигнала.

На основе генов ядра, оболочки и облака, пан-геномы можно разделить на два типа— открытые и закрытые (рисунок 1.8). Закрытыми принято считать пан-геномы, которые быстро выходят на плато по числу генов; иными словами, секвенирование геномов новых представителей исследуемого вида не добавляет новых генов. Открытые пан-геномы, напротив, увеличиваются при добавлении каждого нового генома.

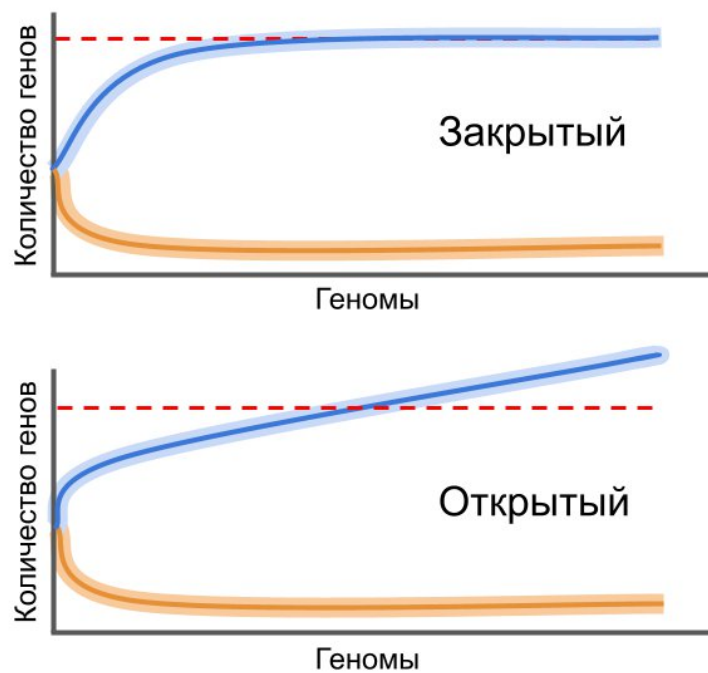


Рисунок 1.8. Тип пан-геномов. Синяя линия показывает сам пан-геном, оранжевая его ядро, а именно гены, присутствующие в геномах всех представителей вида.

Тип пан-генома отражает особенности генетической изменчивости вида. Открытый пан-геном чаще встречается у микроорганизмов (Hou и Pang, 2024), в частности бактерий, для которых обычно характерны крупные размеры популяций, высокая частота мутаций и интенсивный горизонтальный перенос генов, позволяющий им приобретать новый генетический материал. Это обеспечивает бактериям значительную экологическую пластичность и способность быстро адаптироваться к меняющимся условиям окружающей среды. Для растений, в особенности для сельскохозяйственных культур, более характерен закрытый тип пан-генома (Hou и Pang, 2024). Это связано с тем, что

у растений более слабо выражен горизонтальный перенос генов, а генетическая изменчивость в большей степени обеспечивается перестройками уже существующего набора генов.

Для селекции и генетических исследований больше подходит закрытый тип пан-генома, т.к. позволяет заранее оценить, сколько линий или сортов достаточно проанализировать, чтобы охватить почти всё разнообразие генов вида. Тогда как открытый тип пан-генома в основном используется для поиска новых, хозяйственно ценных вариантов (Hou и Pang, 2024).

1.9.2 Пан-транскриптомы

Еще одним из методов анализа геномного состава у нескольких представителей какого-либо таксона является анализ его транскриптомов. Нуклеотидные последовательности транскриптов позволяют оценить присутствие генов в геноме только в том случае, если они экспрессируются в какой-либо ткани или органе растения. Таким образом, определение пан-транскриптома можно сформулировать так: пан-транскриптом отражает совокупность всех молекул РНК, присутствующих в данном виде.

Применение пан-транскриптома позволяет определить, какие гены и их изоформы экспрессируются в конкретных тканях или при определённых условиях (Zhong et al., 2024). Сравнение транскриптомов особей из различных экологических ниш даёт возможность связать вариации экспрессии генов (ePAV) с адаптивными признаками и выявить локусы, которые могли бы остаться незамеченными при анализе только геномных последовательностей (Hirsch et al., 2014). Помимо биологических преимуществ, использование пан-транскриптома экономически более выгодно, когда требуется проанализировать сотни представителей вида: он опирается на массовые RNA-seq данные, которые можно получить быстро и относительно дёшево даже для сотен генотипов (Zhong et al., 2024; Hirsch et al., 2014).

Jin et al. (Jin et al., 2016) провели анализ пан-транскриптома 368 инбредных линий кукурузы. Анализ позволил выявить более двух тысяч последовательностей, которые не были представлены в референсном геноме

кукурузы, среди них — гены, ответственные за ответ на биотический стресс. Были проанализированы вариации, ассоциированные с уровнем экспрессии генов (eQTL). Результаты анализа были спроецированы на метаболические сети, что позволило уточнить механизмы их функционирования.

В работе Ма (Ma et al., 2019) проанализировано 288 экспериментов по секвенированию транскриптома ячменя. Среди собранных транскриптов около 30 % не обнаружили сходства с референсным геномом. Результаты анализа пан-транскриптома показали, что гены устойчивости к патогенам более многочисленны у дикорастущего ячменя; такие гены в процессе domestikации были подвержены более сильному давлению отбора по сравнению с генами других представителей вида.

1.9.3 днРНК в пан-геномах и пан-транскриптомах

Несмотря на растущее количество работ по изучению пан-геномов и пан-транскриптомов, большинство из них было сосредоточено на белок кодирующих генах. Исследование днРНК в этом контексте ограничивалось их классификацией и анализом экспрессии. Так, в работе (Jin et al., 2016) при анализе пан-транскриптома кукурузы (*Z. mays*) было выявлено 644 новых днРНК. В работе (Zhong et al., 2024), посвященной анализу пан-транскриптома риса, были выявлены 3105 днРНК, 2182 из которых были охарактеризованы как новые. Однако в обеих работах дальнейший анализ этих днРНК не проводился. При исследовании пан-транскриптома люцерны (*Medicago sativa* L.) в тканях листьев, стебля и корней было выявлено 11 677 новых днРНК (Medina, 2021). Также данный анализ позволил идентифицировать 2105 высокоэкспрессирующихся и 1894 низкоэкспрессирующихся днРНК.

Более углубленное изучение днРНК было проведено (Kornienko et al., 2024). Авторы исследовали распространенность этих генов в природных популяциях и причины вариабельности их активности между растениями. Для этого они использовали 499 образцов *A. thaliana*. В результате было идентифицировано 11 295 локусов днРНК, составивших, по сути, пан-транскриптом днРНК, однако сам этот термин авторы не использовали. Из

полученных локусов авторы выделили 10 315 (91%) транскриптов, которые были отнесены к новым днРНК. Классификация показала, что антисмысловые днРНК составляли 72 % от общего набора, межгенные - 20 %. Также авторы выделили отдельный класс днРНК, антисмысловых по отношению к мобильным элементам, на который приходилось 6 % от общего набора. Исследование выявило, что в отличие от белок кодирующих генов, активных практически во всех образцах, большинство днРНК экспрессировались менее чем в 5 % образцов. Межгенные днРНК демонстрировали наибольшую межпопуляционную изменчивость, сопоставимую с транспозонами. Также авторы установили, что вариабельность экспрессии днРНК у *A. thaliana* определялась эпигенетическими паттернами.

Таким образом, применение пан-геномов и пан-транскриптомов повысило эффективность выявления днРНК и увеличило их количество по сравнению с анализом генома одного представителя вида. Вместе с тем работы, посвященные детальному изучению структуры, функций и эволюции днРНК на уровне пан-транскриптомов, остаются немногочисленными. Следует отметить, что в них не рассматриваются такие признаки набора днРНК, как открытость/закрытость пан-транскриптома, классификация по представленности в ядре или оболочке пан-транскриптома, а также не проводится систематическая оценка скоростей замен по сравнению с мРНК.

1.9.4 Биоинформатические методы анализа пан-геномов и пан-транскриптомов

В биоинформатическом анализе пан-генома можно выделить следующие основные этапы:

1. Сборка последовательностей пан-генома;
2. Выделение консервативных и вариабельных участков геномных последовательностей;
3. Идентификация/предсказание и функциональная аннотация генов;
4. Идентификация полиморфизмов;
5. Хранение, обеспечение быстрого доступа и визуализация пан-геномных данных.

Для сборки пан-геномов существуют следующие стратегии: сборка-выравнивание; метагеномный подход; картирование-сборка (Golicz et al., 2016; Hurgobin и Edwards, 2017; Rouard и Sabot, 2019).

Сборка-выравнивание. Эта стратегия заключается в *de novo* сборке последовательностей каждого представителя таксона по отдельности с последующим выравниванием последовательностей друг на друга, а также относительно референсного генома, чтобы уменьшить избыточность и определить набор общих и вариабельных участков последовательностей. Для сборки генома разработано несколько программных пакетов: Velvet (Zerbino и Birney, 2008), SOAPdenovo (Xie et al., 2014), ALLPATHS (Butler et al., 2008) и MaSuRCA (Zimin et al., 2013). Этот подход требует много времени и вычислительных ресурсов. Стратегия сборки *de novo* использовалась для анализа пан-генома сои (Li et al., 2014), капусты (Golicz et al., 2016), люцерны (Shannon и Samac, 2024).

Метагеномный подход. Эта стратегия заключается в объединении всех секвенированных фрагментов из разных представителей таксона в один пул и *de novo* сборке последовательностей пан-генома из этих фрагментов. После этого каждый собранный контиг относится к определенному геному путем выравнивания исходных прочтений этого представителя на метагеномную сборку и последующей оценки покрытия контигов. Данный метод позволяет работать с результатами секвенирования с низким уровнем покрытия. Метагеномный подход использовался для анализа геномов риса (Yao et al., 2015), томата (Gao et al., 2019).

Выравнивание-сборка. Эта стратегия использует сборку одного полного генома (референсной последовательности) в качестве основы для сборки геномов остальных представителей таксона (*guide assembly*). Прочтения из одного представителя вида выравниваются относительно референсного генома; прочтения, которые не совпали, отсеиваются и собираются отдельно. Референсный геном обновляется новыми собранными последовательностями, далее образцы сравниваются с данным референсным геномом. Данный метод позволяет сократить время построения пан-генома. В случае если геномный

сегмент существует сразу в нескольких образцах, данный сегмент будет интегрирован из первого образца, в то время как метод *de novo* создает два полных генома. Такая стратегия использовалась при анализе пан-генома подсолнечника (Hübner et al., 2019).

Следует также отметить, что в ряде исследований авторы не использовали сборку геномных последовательностей, а выравнивали короткие прочтения на референсный геном. Такой подход позволяет оценить однонуклеотидные вариации, выявить связь этих полиморфизмов с фенотипическими характеристиками растений. Также существуют методы, которые на основе выравнивания коротких прочтений позволяют оценить структурные перестройки, дубликации и потери генов (Zhao et al., 2013). Метод выравнивания использовался при анализе пан-транскриптома кукурузы (Hirsch et al., 2014), при оценке изменения числа копий генов при анализе пан-генома картофеля (Żmieńko et al., 2014).

Аннотация генома позволяет идентифицировать последовательности генов в геномах представителей таксона, на основе сравнения их последовательностей определить ортологичные гены, а также семейства генов ядра, оболочки и облака. Для автоматической аннотации пан-геномов разработан ряд программных пакетов, выполненных в виде вычислительных конвейеров. Эти инструменты охватывают все основные этапы пан-геномного анализа — от идентификации генов до функциональной интерпретации и визуализации результатов.

Программа PGAP (Zhao et al., 2012) осуществляет масштабный поиск генов, проводит функциональную аннотацию, обогащение кластеров ортологичных генов терминами генной онтологии, анализ эволюции видов, производит анализ генного состава пан-генома, а также разделение генов по группам ядра, оболочки и облака в пан-геноме. В обновленной версии этой программы PGAP-X (Zhao et al., 2018) дальнейшее развитие получили методы представления и визуализации результатов анализа пан-геномов.

Пакет программ PpsPCP (Tahir Ul Qamar et al., 2019) разработан для идентификации вариаций наличия/отсутствия генов (PAV) в пан-геномах. Анализ основан на полногеномном сравнении последовательностей представителей

таксона и референсного генома в несколько раундов с последовательной коррекцией как набора генов, так и участков их выравнивания в референсном геноме. В итоге создается набор генов пан-генома путем объединения последовательностей отдельных геномов с референсным геномом и их аннотации.

Программа BPGA (Chaudhari, Gupta и Dutta, 2016) реализует широкие возможности по анализу пан-геномов: кластеризацию генов на основе сходства последовательностей, анализ наличия/отсутствия ортологов, построение графика зависимости размеров пан-генома и его универсальной части в зависимости от количества геномов, реконструкцию филогенетического дерева между представителями таксона, анализ метаболических путей и функциональной аннотации, оценку отклонений GC состава, расчет различных статистических характеристик пан-генома и ряд других возможностей.

Программа panX (Ding, Baumdicker и Neher, 2018) направлена на идентификацию кластеров ортологичных генов. Для этого используется кластеризация на основе сравнения последовательностей, верификация и уточнение состава кластеров на основе анализа эволюционных расстояний и филогенетической реконструкции. Это позволяет оценить ассоциацию между геномным составом индивидуальных представителей таксона и их фенотипов.

Программа Pan4Draft (Veras et al., 2018) разработана для получения улучшенной аннотации пан-геномов за счет добавления к ней информации о последовательностях незавершенных геномов (unfinished genomes). Это геномы, у которых аннотация и сборка до уровня хромосом не завершена, но их последовательности содержат фрагменты геномной ДНК и представляют ценную информацию о разнообразии геномов вида.

Следует отметить, что все рассмотренные пакеты реализуют многоэтапную обработку больших данных, включая в анализ различные методы и алгоритмы, которые отчасти реализованы отдельными пакетами. Такие программы удобно реализовывать в виде вычислительных конвейеров.

1.10 Конвейеры обработки данных

На сегодняшний день за счет применения технологий секвенирования нового поколения стремительно увеличивается количество экспериментов по массовому секвенированию РНК (RNA-seq). Работа с пан-геномами и пан-транскриптомами также предполагает обработку больших объемов данных. Чтобы повысить эффективность анализа транскриптомов для большого количества данных, активно разрабатываются вычислительные конвейеры, которые позволяют выполнять стандартизированные этапы обработки в режиме, не требующем контроля пользователем вплоть до получения конечного результата. Главными достоинствами существующих конвейеров являются возможность сборки транскриптома для исследуемого организма и анализ дифференциальной экспрессии днРНК. Более подробное сравнение функционала биоинформатических конвейеров представлено в таблице П.2.

Биоинформатические конвейеры позволяют автоматизировать и ускорить этапы анализа днРНК. На сегодняшний день выделяют несколько архитектур построения таких конвейеров. Snakemake (Köster и Rahmann, 2012) представляет собой систему управления рабочими процессами, основанную на языке Python. Основным достоинством Snakemake является синтаксис, приближенный к псевдокоду и не требующий глубоких знаний программирования, что делает систему доступной для широкого круга пользователей. Nextflow (Di Tommaso et al., 2017) - система управления рабочими процессами, написанная на языках Groovy и Java и разработанная для биоинформатики и научных вычислений. В основе Nextflow лежит парадигма программирования потоков данных. В этой модели задачи выполняются автоматически, как только становятся доступны необходимые входные данные, что обеспечивает более гибкое и эффективное параллельное выполнение. Также существуют веб-платформы, такие как Galaxy (Jalili et al., 2020), для доступного и воспроизводимого проведения биоинформатических анализов. В отличие от большинства других систем управления рабочими процессами, Galaxy предоставляет полноценный графический интерфейс пользователя (GUI), который позволяет создавать и выполнять рабочие процессы без написания кода.

Однако, несмотря на достоинства конвейеров, они имеют свои недостатки. Большинство из них разработаны на основе транскриптомных данных модельных организмов (человека, мыши, плодовой мушки). Вследствие этого для исследования другого организма нужно либо использовать уже имеющиеся модели, что может внести ошибки в предсказание днРНК, либо дополнительно обучать модель предсказания для данного организма, а затем вносить эти данные в код конвейера, что также может привести к неправильной работе программы.

Отметим, что, несмотря на общие черты, между структурой и функциями днРНК растений и животных существуют отличия (Golicz, Singh и Bhalla, 2018). Например, в работе (Costa Negri, 2020), посвященной анализу программ предсказания днРНК, было показано, что программы, разработанные для человека, плохо предсказывают днРНК растений и наоборот, что свидетельствует о различиях в структурных особенностях днРНК между человеком и растениями. Однако экспериментальных работ, направленных на сравнение структуры днРНК растений и животных, пока нет. Поэтому для более точного выявления днРНК необходимы дополнительные фильтры последовательностей, которые предоставляют не все конвейеры.

Таким образом, выбор системы управления рабочими процессами зависит от конкретных требований проекта, навыков пользователя и вычислительной инфраструктуры. Snakemake предлагает хороший баланс между простотой использования и функциональностью для пользователей, предпочитающих работать с языком Python. Nextflow обеспечивает масштабируемость и имеет доступ к готовым биоинформатическим конвейерам через собственную базу данных. Однако независимо от выбранной системы, использование данного подхода значительно упрощает работу биоинформатика.

1.11 Кукуруза и экономическая значимость

Кукуруза (*Z. mays*) - вид рода Кукуруза (*Zea*), принадлежащий к трибе Андропогоновые (*Andropogoneae*) семейства Злаки (*Poaceae*). Кукуруза является однолетним травянистым растением, имеющим соцветие типа початок (у женских цветков) и метёлку (у мужских цветков). Плод кукурузы - зерновка. Для вида кукурузы *Z. mays* характерно перекрёстное опыление с помощью ветра.

На протяжении многих лет отношения внутри рода *Zea* были предметом жарких споров (Buckler et al., 2006). Основная сложность в таксономии кукурузы и определении ее ближайших родственников заключалась в отсутствии початковидного соцветия у всех известных растений, кроме кукурузы (рисунок 1.9А) (Doebley et al., 1990). Несмотря на растущее признание того, что теосинт (*Z. mays*) является непосредственным предком кукурузы (*Z. mays ssp. mays*), консенсус был достигнут не сразу. Было выдвинуто две ведущие гипотезы. В конце 1930-х годов была предложена «трёхчастная гипотеза» (Mangelsdorf, 1938; 1974). Эта гипотеза утверждала, что кукуруза была одомашнена от какого-то неизвестного дикого предка, вероятно, растения со структурами, напоминающими современный кукурузный початок. Уточнялось, что дикий предок кукурузы из Южной Америки либо вымер, либо не обнаружен. Вторая - «гипотеза теосинта» - была выдвинута в 1939 году Бидлом (Beadle, 1939). Как можно понять из названия, данная гипотеза предполагает, что кукуруза - это просто одомашненная форма теосинта. Бидл полагал, что благодаря искусственному отбору, проводившемуся древними популяциями, несколько небольших мутаций с большими эффектами могли превратить теосинт в кукурузу. Однако с развитием молекулярно-генетических технологий удалось подтвердить гипотезу Бидла (Doebley et al., 1984; Matsuoka et al., 2002). Анализ микросателлитных локусов показал, что теосинт (*Z. mays ssp. parviglumis*) из долины реки Бальсас является предком современной кукурузы (Matsuoka et al., 2002).

На сегодняшний день кукуруза — одна из ведущих зерновых культур мирового земледелия. Кукуруза на зерно возделывается в мире на площади более 140 млн. га (Shindin, 2009). Широкому распространению и увеличению производства кукурузы способствуют высокая урожайность и широкие возможности использования как продукта питания, так и ценного корма для сельскохозяйственных животных (Korobka et al., 2015; Shindin, 2009). В Российской Федерации кукуруза является одной из важнейших зерновых культур и имеет высокую экономическую, экологическую и социальную значимость в аграрном секторе (Korobka et al., 2015; Shpaar et al., 2006), рисунок 1.9Б.

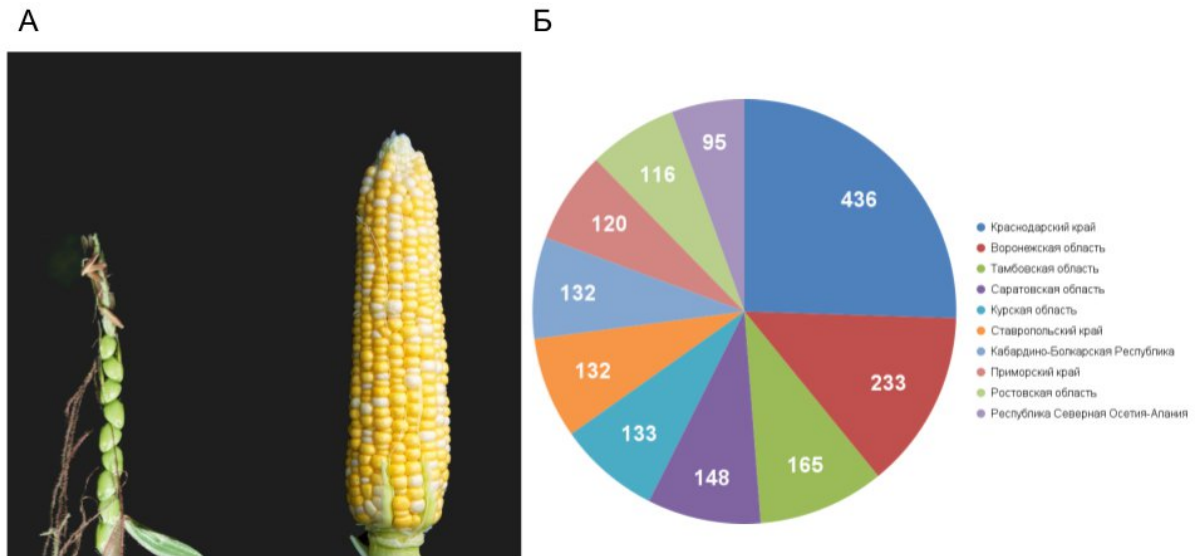


Рисунок 1.9. А) - слева направо: теосинт (*Z. mays ssp. parviglumis*), кукуруза (*Z. mays ssp. mays*); изображение взято из базы данных Klau (<https://klau.club/57960-predok-kukuruzy.html>). Б) - площади, на которых в разных странах высевалась кукуруза в 2025 году, тыс. га (<https://xn--e1alid.xn--p1ai/journal/publication/posevnye-ploshchadi-pod-kukuruzoiy-v-2025-godu-sostavyat-okolo-262-mln-gektarov>).

Калорийность кукурузы, обладающей ценной кормовой, пищевой, а также лечебной значимостью, составляет 325 ккал на 100 г. Основными составляющими являются моно-, ди- и полисахариды, общее содержание которых достигает около 60 %. Кроме того, кукуруза богата ненасыщенными жирными кислотами, витаминами группы В, макро- (калий, фосфор, магний и др.) и микроэлементами (алюминий, бор, медь и др.). Кормовая ценность кукурузы составляет 1.34 кормовых единиц (КЕ) (Сотченко et al., 2009). Зерно кукурузы используют для производства масла, широко применяемого в пищевых целях, при изготовлении кондитерских изделий и других продуктов пищевого назначения (в том числе в блюдах национальной кухни разных стран). Промышленность перерабатывает не только зерно, но и стержни, стебли, обертки початков, получая из них жидкую смолу, бутиловый спирт, фурфурол, изоляционные прокладки, линолеум, краски, клей, медикаменты. По данным ФАО, в настоящее время во всем мире из

кукурузы производят более 500 различных основных и побочных продуктов (Полетаев et al., 2008; Стулин et al., 2009).

На сегодняшний день кукуруза имеет один из наиболее полно собранных геномов. Его размер составляет 2.1-2.2 Гб, он насчитывает 10 хромосом. Полнота генома по критерию BUSCO составляет около 96 %. Геном содержит примерно 39 000 - 40 000 белок кодирующих генов (включая 39 756 генов согласно RefGen_v5), а также 20 163 предполагаемых длинных некодирующих РНК. Геном кукурузы характеризуется высокой долей мобильных элементов: 65-85 % генома составляют мобильные элементы и повторяющиеся последовательности.

Для кукурузы уже проведены работы по сборке пан-геномов и пан-транскриптомов. Исследование пан-транскриптома 503 инбредных линий кукурузы (Hirsch et al., 2014) позволило выявить генетическое разнообразие в белок кодирующих генах. Авторы показали структуру пан-транскриптома, где на консервативные гены (ядро) приходится 16 % транскриптов, а на варибельную часть (оболочка + облако) — 84 % транскриптов. Также в работе обнаружено более полутора миллионов однонуклеотидных вариаций, выявлены мутации, ассоциированные с признаками развития растений (время прохождения ряда фаз роста). В работе (Jin et al., 2016) проведен анализ 368 инбредных линий кукурузы. Авторы охарактеризовали вариации экспрессии генов (ePAV) и их регуляцию, а также выявили более двух тысяч новых последовательностей мРНК. На основе полученных вариаций генов был проведен анализ ассоциаций (GWAS) для агрономических и метаболических признаков, что позволило обнаружить влияние варибельных мРНК на гетерозис. В работе (Hufford et al., 2021) выполнена сборка пан-генома из 26 линий кукурузы, являющихся представителями NAM-популяции (Nested Association Mapping). Данные 26 образцов, по утверждению авторов, наилучшим образом представляют генетическое разнообразие кукурузы. В результате выявлено 103 тысячи генов, 50 % из которых принадлежат консервативной части. Было показано, что консервативные гены обладают высокой синтенией с близкородственными видами (более 56 %), тогда как варибельные гены - лишь около 2 % синтенности. Функции консервативных генов в основном связаны с базовыми и критическими

клеточными процессами, что согласуется с литературными данными о данном типе генов в пан-геноме. Gui с соавторами в недавней работе (Gui et al., 2022) провели сборку пан-генома кукурузы, включающего 721 линию. Авторы ставили целью создание комплексной карты генетических вариаций для селекции кукурузы. Это позволило учесть генетическое разнообразие диких родственников кукурузы для повышения стрессоустойчивости и урожайности. Также были идентифицированы гены-кандидаты, отсутствующие в референсном геноме кукурузы. В данной работе было показано, что пан-геном кукурузы является закрытым и что для получения набора из 99 % всех белок кодирующих генов пан-генома достаточно уже 93 образцов кукурузы.

Анализ днРНК кукурузы до настоящего времени был в основном направлен на исследование их участия в регуляции абиотического стресса (Zhang et al., 2014; Hu et al., 2022; Waititu et al., 2022). В работе (Zhang et al., 2014) выявлено 664 днРНК, которые демонстрируют дифференциальную экспрессию в листьях и корнях в ответ на засуху, что свидетельствует об их потенциальной роли в регуляции адаптации к водному стрессу. Исследование (Hu et al., 2022) выявило 993 днРНК, дифференциально экспрессирующихся в ответ на засуху. Аннотация связанных с выявленными днРНК белок кодирующих генов показала, что днРНК также могут участвовать в регуляции фотосинтетических процессов. Более масштабные исследования были проведены в работе (Han et al., 2019). Авторы проанализировали транскриптомные данные для 1027 библиотек RNA-seq кукурузы с целью лучше понять эволюционную роль днРНК в процессе одомашнивания кукурузы и её адаптации к сельскохозяйственным условиям. В результате анализа было идентифицировано 18 165 днРНК. Полученные днРНК были разделены на два класса: поливидовыми считались днРНК, имеющие не менее 50 % идентичности с генами теосинта, а внутривидовые — те, которые такого сходства не имеют. В работе показано, что днРНК экспрессируются в целом тканеспецифично, при этом внутривидовые днРНК с большей вероятностью обнаруживались в репродуктивных органах (например, в пыльниках), чем поливидовые днРНК.

Таким образом, несмотря на значительный прогресс в секвенировании генома кукурузы и в исследованиях пан-геномов, изучение роли днРНК ограничивалось лишь их идентификацией, анализом первичной структуры и экспрессии в рамках одного представителя вида. Тогда как работы, направленные на анализ разнообразия, характеристик первичной структуры, экспрессии и эволюции днРНК в рамках единой концепции пан-транскриптома для кукурузы, до сих пор не проводились. Подобные исследования позволили бы раскрыть молекулярные механизмы адаптации растения к изменению природных условий, а также создать новые подходы к улучшению урожайности и стрессоустойчивости сортов.

1.12 Заключение по обзору литературы и постановка задачи исследования

В обзоре литературы рассмотрена основная информация о длинных некодирующих РНК: их структура, молекулярные и биологические функции, эволюция, а также методы секвенирования и биоинформатического анализа. Показана важность и незаменимость днРНК во многих клеточных процессах, таких как активация и подавление экспрессии генов, импринтинг и деметилирование ДНК, РНК-интерференция, ремоделирование хроматина и другие. У растений днРНК участвуют в регуляции устойчивости к различным видам стресса, принимают участие в развитии плодов, корней и листьев. днРНК выполняют функции как генетических, так и эпигенетических регуляторов экспрессии генов. Они могут действовать как цис-регуляторные элементы, взаимодействуя с соседними генами, или как транс-регуляторные элементы, оказывая влияние на гены, расположенные на удалении от места их синтеза. Это разнообразие функций делает днРНК важным объектом для понимания молекулярных механизмов роста и развития растений, их ответа на биотический и абиотический стресс.

Однако, несмотря на растущее число исследований, посвященных структурно-функциональному анализу днРНК, данный тип молекул по-прежнему остается малоизученным, особенно у растений. Это связано со множеством факторов, которые необходимо учитывать при идентификации днРНК. Прежде

всего, размер днРНК схож с размером некоторых белок кодирующих генов, что приводит к ложному определению этих молекул. Кроме того, исследования показывают, что последовательности днРНК претерпевают быструю эволюцию, закономерности которой пока непонятны. Некоторые днРНК подвержены высокой скорости мутаций, и в результате их нуклеотидные последовательности могут значительно различаться даже у близкородственных организмов. Таким образом, необходимо расширять знания об этих молекулах. Однако использование экспериментальных методов в масштабах всего генома затруднено вследствие их высокой трудоемкости и стоимости.

Применение биоинформатических методов позволяет проводить комплексный анализ днРНК, включающий предсказание, фильтрацию и аннотацию. Методы сравнительного эволюционного анализа позволяют изучать закономерности изменчивости данных молекул. Классификация днРНК позволяет лучше идентифицировать данные молекулы. Однако, как было указано выше, днРНК демонстрируют высокую внутривидовую специфичность, поэтому выборку данных молекул необходимо расширять за счет исследования большего количества представителей одного вида.

Для этих целей используется концепция пан-генома, которая предполагает охват последовательностей белок кодирующих генов, подверженных вариациям наличия/отсутствия (presence-absence variations, PAV) у нескольких представителей вида и, возможно, отсутствующих в референсной последовательности. Пан-транскриптомные методы позволяют выделять консервативные и вариабельные части генов в транскриптомах, однако для днРНК работ такого масштаба до настоящего времени практически не было. Также работа с пан-транскриптомами требует обработки больших объемов данных из множества индивидуальных транскриптомов, что делает ручной анализ долгим и трудозатратным процессом.

В связи с этим возникает задача разработки автоматизированных конвейеров, позволяющих производить массовый анализ днРНК в транскриптомах, их качественное предсказание и аннотацию, а также сборку пан-транскриптомов с разделением на кодирующие и не кодирующие части.

Разработка подобных конвейеров позволит выявить редкие аллели и генетические вариации, которые присущи лишь некоторым представителям исследуемого вида. Применение методов глубокого обучения для предсказания днРНК в контексте пан-транскриптомного анализа расширит знания о характеристиках различных типов днРНК и будет способствовать развитию методов аннотации. Сравнительный анализ пан-транскриптомов позволит найти общие белок кодирующие гены и днРНК между исследуемыми организмами, что расширит знания об эволюции как кодирующих, так и не кодирующих компонентов генома.

В качестве изучаемого организма нами была выбрана кукуруза, которая является одной из ведущих зерновых культур мирового земледелия. Кукуруза обладает одним из наиболее полно секвенированных референсных геномов, который можно рассматривать наравне с геномами модельных организмов. Кроме того, для кукурузы уже проведены работы по сборке пан-геномов и пан-транскриптомов, а в литературе известно большое количество аннотированных днРНК. Это делает кукурузу идеальным объектом для разработки и апробации методов анализа днРНК в масштабах пан-транскриптома, требующих обработки значительных объемов данных.

ГЛАВА 2 Материалы и методы

2.1 Разработка конвейера ICApnoLncRNA для идентификации и анализа днРНК в транскриптомах растений

Для выявления последовательностей днРНК в транскриптомах растений нами был разработан вычислительный конвейер ICApnoLncRNA, схема которого представлена на рисунке 2.1.

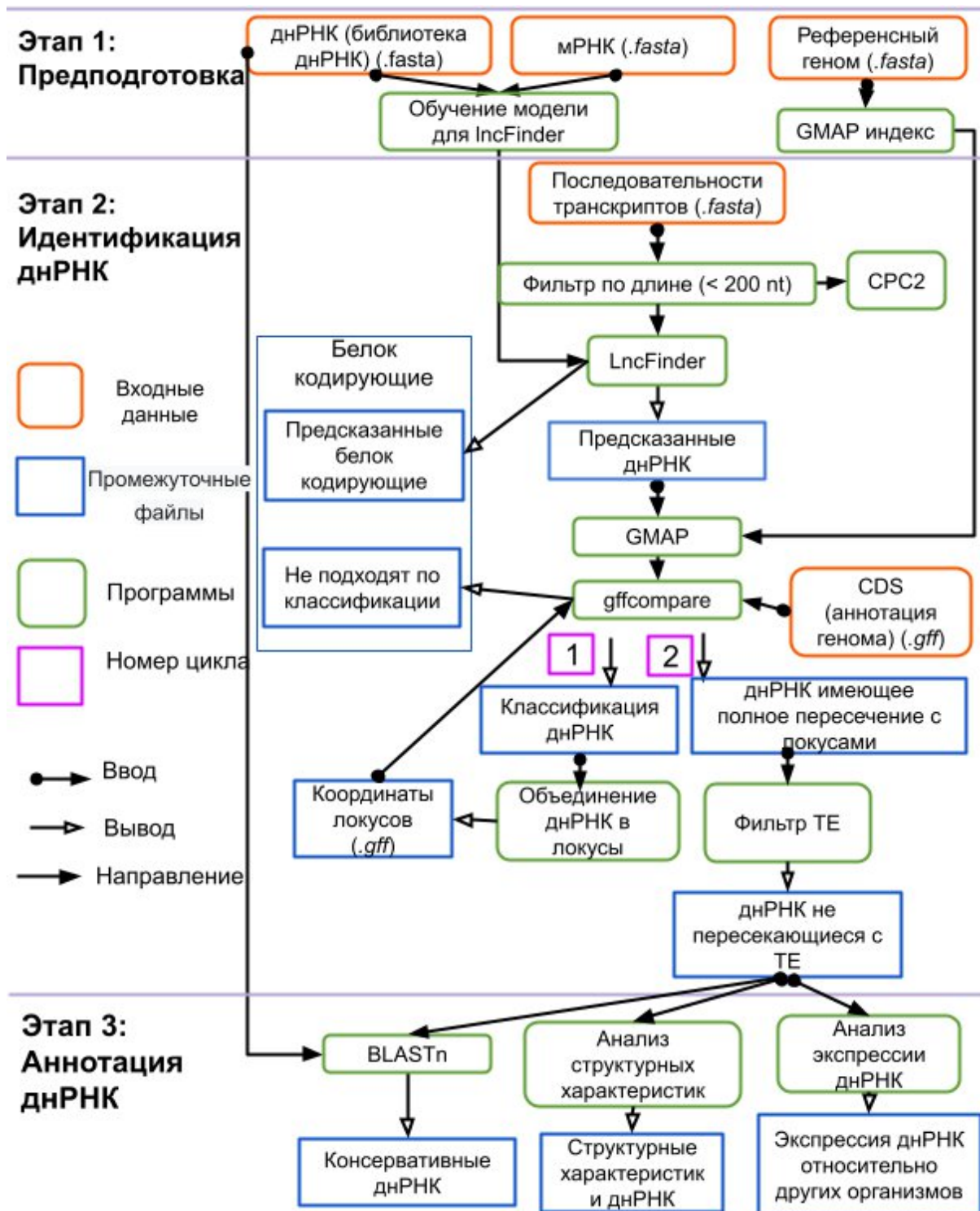


Рисунок 2.1. Схема биоинформатического конвейера ICannoLncRNA. В оранжевых прямоугольниках показаны входные данные, в синих - исходные данные после определенного этапа конвейера, в зеленых - используемая программа, фиолетовыми прямоугольниками обозначен выбор цикла: (1) при наличии информации о мобильных элементах, (2) при отсутствии.

Анализ включает несколько основных блоков: предварительная обработка данных (этап 1, рисунок 2.1), идентификация днРНК (этап 2, рисунок 2.1), аннотация днРНК (этап 3, рисунок 2.1). Эти блоки детально будут описаны ниже.

2.1.1 Предварительная обработка данных

На первом этапе работы конвейера происходит построение индексного файла для референсной геномной последовательности с помощью программы GMAP (Wu, 2005) и обучение модели распознавания днРНК.

Для точного разделения выборки РНК на кодирующие белки и некодирующие применяют различные биоинформатические методы. К ним относятся как классические методы, основанные на анализе последовательностей и структурных особенностях пространственной организации белков, так и методы машинного и глубокого обучения.

В данном конвейере использовалась программа LncFinder v1.1.4 Han et al., 2019, реализующая подход машинного обучения для классификации транскриптов. LncFinder позволяет обучать модель распознавания днРНК на основе известных последовательностей днРНК и мРНК специфического генома с известной аннотацией. Эта программа использует алгоритм машинного обучения для классификации последовательностей и три вида параметров: частоты гексамеров в нуклеотидных последовательностях, физико-химические свойства нуклеотидов и особенности предсказанной для последовательности вторичной структуры. Последовательности для обучения модели включают полный набор известных мРНК и днРНК для исследуемого генома и предоставляются в виде двух соответствующих файлов в FASTA-формате. В результате работы конвейера из этих полных данных формируются обучающая и тестовая выборки для программы LncFinder. С учетом того, что количество известных днРНК в аннотированных геномах, как правило, существенно меньше, чем мРНК (например, для исследуемого генома кукурузы версии 40 аннотировано 33 725 мРНК и 2535 днРНК), для анализа используются все известные последовательности днРНК, а количество последовательностей мРНК берется в два раза большим, чем днРНК. При этом указанное количество мРНК

выбирается из общей выборки независимым случайным образом (без возвращения). Такое соотношение количества мРНК и днРНК в обучающих данных было использовано в работе Nan et al., 2019 и позволяет исключить сильное смещение обучающей выборки в сторону перепредставленности последовательностей мРНК.

В процессе обучения сети сформированные таким образом данные - как мРНК, так и днРНК - делятся случайным образом в пропорции 80% для обучающей выборки и 20% для тестовой. С использованием обучающих данных производится оптимизация параметров модели, а на тестовых данных - оценка её точности. Точность оценивается с помощью метрики $F1$ (Sasaki, 2007):

$$F1 = \frac{2TP}{2TP + FP + FN}$$

где, TP - предсказанные днРНК соответствуют известным днРНК, FP - предсказанные днРНК соответствуют известным мРНК, TN - предсказанные мРНК соответствуют известным мРНК, FN - предсказанные мРНК соответствуют известным днРНК. $F1$ -мера представляет собой гармоническое среднее между точностью и полнотой, чем ее значение ближе к 1, тем выше точность предсказания днРНК.

Дополнительно для оценки точности модели используется коэффициент корреляции Мэтьюса (MCC , *Matthews Correlation Coefficient*) (Matthews, 1975). MCC учитывает все четыре категории матрицы ошибок: истинно положительные (TP), истинно отрицательные (TN), ложно положительные (FP) и ложно отрицательные (FN) результаты. Это делает его надежной и сбалансированной метрикой, особенно при работе с несбалансированными выборками данных, что характерно для задач предсказания днРНК, где количество не кодирующих последовательностей существенно меньше кодирующих. Метрика вычисляется по следующей формуле:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

где, TP - количество правильно классифицированных днРНК, TN - количество правильно классифицированных мРНК, FP - количество мРНК, ошибочно предсказанных как днРНК, FN - количество днРНК, ошибочно предсказанных как мРНК.

Значение коэффициента MCC варьируется в диапазоне от -1 до +1. Чем значение метрики ближе к +1, тем выше вероятность того, что все последовательности распределены по классам верно. Значение 0 свидетельствует о том, что предсказания модели не лучше случайного угадывания. Значение -1 означает полное несоответствие между предсказанными и фактическими классами.

Процедура обучения и тестирования, описанная выше, производится посредством 5-кратной кросс-валидации. При этом отбор последовательностей в тестовые и обучающие выборки на каждой итерации производится независимым случайным образом, так что последовательности тестовой выборки не встречаются в обучающей. Для оценки устойчивости подбора оптимальных параметров модели в дополнение к экспериментам с 5-кратной кросс-валидацией мы провели сравнение распределений мер $F1$ и MCC для выборки с 10-кратной кросс-валидацией. Распределения параметров точности предсказания днРНК на этих двух выборках сравнивали с использованием критерия Манна - Уитни.

При работе конвейера результаты тестирования модели для каждой из реплик в процессе кросс-валидации приводятся в отдельном файле в CSV-формате. В качестве оптимальных выбираются параметры модели для той реплики, в которой было получено максимальное значение $F1$ и MCC (значения данных метрик коррелируют). Они сохраняются в отдельном файле для дальнейшего анализа.

Наш конвейер позволяет проводить расчёт в двух режимах. Первый режим осуществляет предсказание днРНК на основе частот гексамеров в нуклеотидных последовательностях и физико-химических свойств нуклеотидов (энергию делокализованных электронов). Во втором режиме добавляются параметры вторичной структуры для распознавания днРНК, как это реализовано в пакете LncFinder. Однако этот режим гораздо более затратный по вычислениям (в

десятки раз) по сравнению с первым из-за трудоемкости расчета вторичной структуры РНК и не рекомендован пользователям в случае обработки больших массивов данных (например, при анализе полного транскриптома).

Отметим, что несмотря на высокую точность использованной модели, полученные с ее помощью результаты требуют дополнительной валидации, которая осуществляется на следующем этапе.

2.1.2 Идентификация днРНК

Данный блок состоит из пяти этапов: (1) предсказание молекул-кандидатов в днРНК из входного набора транскриптов с помощью метода LncFinder с параметрами, полученными на предыдущем этапе; (2) картирование последовательностей-кандидатов днРНК на референсный геном; (3) классификация днРНК по локализации в геноме; (4) фильтр транскриптов, являющихся потенциальными ошибками сборки; (5) фильтр транскриптов, пересекающихся с мобильными элементами.

Предсказание кандидатов в днРНК

Для распознавания днРНК в последовательностях транскриптов используется программа LncFinder с параметрами, полученными на предыдущем этапе (см. выше). Параллельно для каждой последовательности с помощью программы CPC2 (Kang, 2017) оцениваются значения следующих параметров: длина открытой рамки считывания, показатель Фиккета, покрытие ORC и величина изоэлектрической точки. Эти параметры сохраняются для каждой последовательности в отдельном файле в качестве дополнительной информации.

В результате работы программы LncFinder транскрипты разбиваются на две выборки: кодирующие и не кодирующие последовательности, идентификаторы последовательностей и их классификация выводятся в CSV-файл. Последовательности, определенные LncFinder как не кодирующие, являются кандидатами в днРНК.

Картирование на референсный геном

Для картирования кандидатов днРНК на референсный геном используется программа GMAP v2020.10.14 (Wu, 2005) с параметрами “--min-intronlength=1 --

```
intronlength=414579 --cross-species --format=gff3_gene --split-large-introns --
npaths=1".
```

Классификация днРНК

Данный шаг необходим для более точного распознавания днРНК. Это связано с тем, что днРНК имеют схожую структуру с белок кодирующими генами (например, длину и экзон-интронную структуру), что может приводить к ложному предсказанию. Однако классификация днРНК на основе их выравнивания относительно белок кодирующих генов дает возможность исключить возможные ошибки при идентификации днРНК за счет специфичных особенностей их выравнивания.

Для того чтобы провести классификацию полученных на предыдущем шаге последовательностей-кандидатов днРНК по их расположению в геноме относительно белок кодирующих генов, применяется программа *gffcompare* v0.11.2 (Pertea, 2020). На выходе каждой последовательности присваивается один из 14 классов, обозначаемых символами и буквами латинского алфавита.

Предварительный анализ показал, что среди картированных на геном последовательностей-кандидатов днРНК в той или иной пропорции представлены все классы из предложенных в работе (Pertea, 2020). Часть последовательностей оказалась полностью совпадающей с последовательностями мРНК (класс '='), что, вероятно, является результатом неточного предсказания программой LncFinder.

Транскрипты, которые выравнивались на гены днРНК, аннотированные в геноме, считаются уже известными днРНК. Они определяются классом *gffcompare* '=' по отношению к генам, представленным в аннотации генома как lincRNA (межгенные днРНК). Подобные последовательности наш конвейер выводит в отдельный файл.

Другая часть последовательностей имеет протяженные выравнивания с экзонами в '+' направлении (классы 'c', 'k', 'j', 'm', 'n', 'o'). Для человека известны экспериментально подтвержденные примеры днРНК, которые совпадают с фрагментами экзонов и считываются в '+' направлении (Sheng, 2018). Для растений такие данные на сегодняшний момент отсутствуют. Кроме

того, было высказано предположение, что транскрипты классов ‘c’, ‘k’, ‘j’, ‘m’, ‘n’, ‘o’ могут представлять собой новые изоформы известных генов, а не днРНК (Halstead, 2021). В силу указанных противоречий в настоящей работе последовательности этих классов не рассматривались как днРНК и исключались из дальнейшего анализа, см. раздел 1.2 Обзора литературы (Karlik, 2019; De Quattro, 2017; Ma, 2013; Chen, 2020). Еще одна часть последовательностей относилась к классам gffcompare ‘e’, ‘s’, ‘p’, ‘u’ (они включали фрагменты интронов и экзонов одновременно). Мы считали, что эти последовательности не являются днРНК, а представляют собой в основном артефакты секвенирования/сборки, и поэтому исключили их из дальнейшего анализа. Таким образом, для уменьшения погрешности в последующем анализе принято решение исключить классы смысловых днРНК из рассмотрения. В итоге рассматривали классификацию днРНК по положению в геноме на три класса (рисунок 2.2):

- интронные - днРНК, расположенные в интроне белок кодирующего гена (класс gffcompare ‘i’).
- антисмысловые - перекрывающиеся с экзонными участками белок кодирующих генов в противоположном направлении (класс gffcompare ‘x’).
- межгенные - днРНК, расположенные между двумя белок кодирующими генами (класс gffcompare ‘u’).

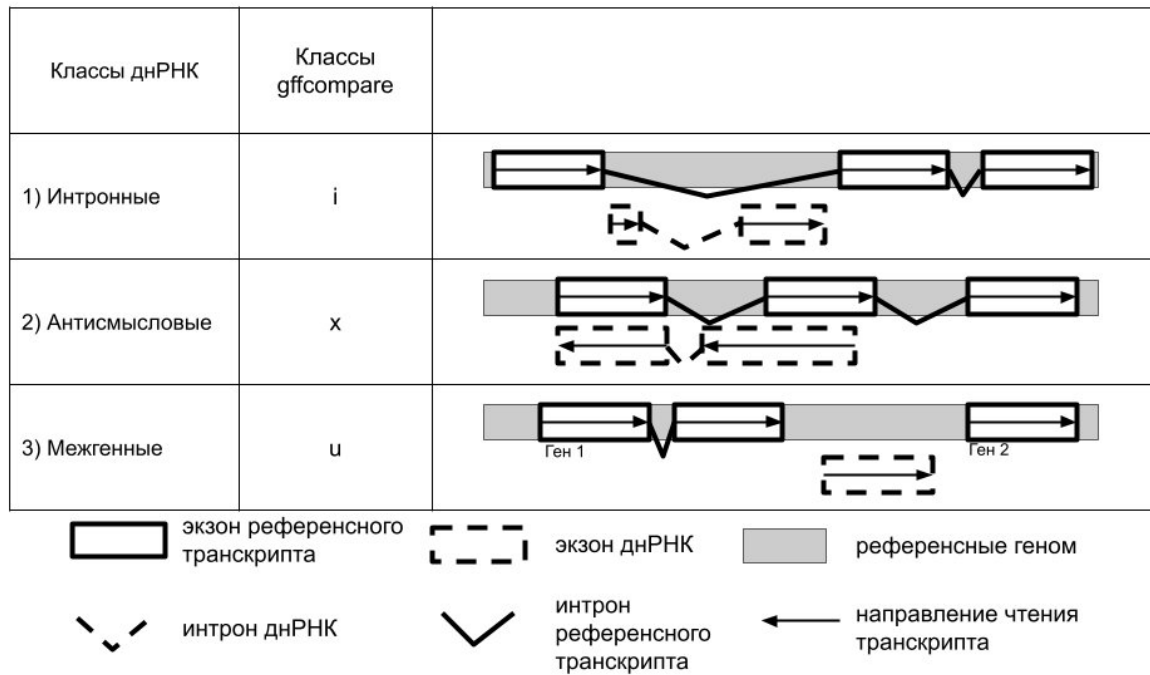


Рисунок 2.2. Идентификация трех классов днРНК по их расположению относительно белок кодирующих генов на основе разметки программы gffcompare. Пунктирными прямоугольниками обозначены днРНК, цельными - экзоны белок кодирующих генов, серым — референсный геном. Интроны днРНК - пунктирная линия, интроны белок кодирующих генов - черная линия. Стрелкой обозначено направление последовательностей.

Таким образом, два класса из трех для выбранной классификации днРНК пересекаются с белок кодирующими генами (классы 'i' и 'x'), а один относится к межгенным локусам (класс 'u').

Фильтрация днРНК

Для последовательностей-кандидатов днРНК, прошедших классификацию (на классы 'i', 'x', 'u') применяется фильтрация коротких последовательностей транскриптов, полученных в ходе *de novo* сборки, которые являются потенциальными ошибками сборки или транскрипционным “шумом”. Для решения этой проблемы транскрипты объединяются в локусы. Геномные локусы днРНК идентифицируются как непрерывный участок хромосомы, с которым пересекается хотя бы один транскрипт днРНК. Сформированный GFF-файл с локусами днРНК и последовательностей-кандидатов днРНК (из которых данные

локусы были сформированы) подаётся на вход программе gffcompare v0.11.2 (Perte, 2020), чтобы выявить транскрипты днРНК, имеющие полное пересечение с локусами (они обозначаются классом “=”). Транскрипты, не имеющие полного пересечения с локусами, удаляются из дальнейшего анализа.

Следующий фильтр, который применяется в данном конвейере, направлен на удаление транскриптов днРНК, имеющих пересечение с мобильными элементами (МЭ). Для применения данной фильтрации пользователю необходимо предоставить файл в формате GFF, содержащий координаты мобильных элементов. В данной работе для получения библиотек и координат мобильных элементов референсного генома был использован пакет EDTA (Ou, 2019). Транскрипты днРНК, которые пересекаются хотя бы с одним нуклеотидом последовательности мобильного элемента, удаляются из дальнейшего анализа.

Последовательности, прошедшие все этапы фильтрации, представляют собой днРНК, имеющие высокую достоверность, и могут быть использованы для дальнейшего анализа.

2.1.3 Выявление видоспецифичности днРНК

Для поиска гомологов с днРНК других организмов конвейер ICAncLncRNA разделяет найденные днРНК на поливидовые и внутривидовые в соответствии с работой (Han et al., 2019). Последовательность днРНК считается поливидовой, если для неё в базах данных известных днРНК обнаруживается гомолог при уровне идентичности последовательностей не менее 50 % у других видов растений. Поиск таких гомологов выполнялся при помощи программы BLASTn v2.9.0+ (Camacho, 2008) со следующими параметрами: “*-outfmt 6 -evaluate 1e-50 -max_target_seqs 1 -perc_identity 50*”. Если для днРНК таких гомологичных последовательностей не обнаруживалось, она считалась внутривидовой, т.е. видоспецифичной.

Для сравнения выявленных нами в транскриптомах днРНК с уже известными последовательностями использовали последовательности из баз данных PNRD (Yi et al., 2015), CANTATAdB v2.0 (Szcześniak et al., 2019), GREENC v1.12 (Gallart et al., 2016), PlncDB v2.0 (Jin et al., 2021), EVLncRNA v2.0 (Zhou et al., 2021). Эти последовательности были объединены в один файл, после

чего полностью идентичные последовательности исключались. В результате было получено 256 091 последовательностей из 16 видов растений, представляющих известные днРНК.

Дополнительно в объединённом наборе известных днРНК и выявленных днРНК, полученных в результате работы конвейера ICAnnoLncRNA, мы выявляли группы сходных последовательностей с помощью программы uCLUST (Prasad, 2015) при уровне идентичности 50%.

2.1.4 Анализ структурных особенностей днРНК и их экспрессии

Анализ структурных особенностей днРНК

Конвейер ICAnnoLncRNA позволяет проводить анализ характеристик днРНК. Рассчитываемые характеристики делятся на две категории: характеризующие структуру днРНК и характеризующие их расположение на хромосомах референсного генома. В первом случае это количество экзонов, распределение их по размерам, распределение размера интронов. Во втором случае это плотность распределения днРНК по хромосомам, распределение антисмысловых днРНК по числу экзонов кодирующих белки генов, с которыми они перекрываются. Также проводится фильтр днРНК, имеющих интрон длиной менее 60 п.н.; конвейер формирует отдельный файл с данными последовательностями, и дальнейший анализ производится без них. Данный фильтр применяется в соответствии с рекомендациями из работы (Cheng, 2018), где авторы ставят под сомнение существование интронов с длиной <60 п.н.

Анализ экспрессии и тканеспецифичности днРНК

Общепринятым считается, что днРНК демонстрируют более низкий уровень экспрессии по сравнению с мРНК, при этом днРНК экспрессируются более специфично для определенных тканей (Derrien et al., 2012). Однако в последнее время начинают появляться работы, демонстрирующие днРНК с уровнями экспрессии, равными таковым для белок кодирующих генов (De Quattro, 2017; Lamin-Samu et al., 2022). Эти данные показывают, что функционально значимые днРНК могут экспрессироваться как на очень низком, так и на сравнимом с мРНК уровне. В связи с этим количественного измерения

экспрессии может быть недостаточно для разделения транскриптов днРНК на функционально значимые и незначимые.

В связи с этим предложенный нами конвейер ICAnnoLncRNA позволяет проводить более углубленный анализ экспрессии, включающий прямое сравнение распределений экспрессии днРНК и мРНК в идентичных тканях, а также оценку тканеспецифичности для эволюционно различающихся групп (поливидовых и внутривидовых) днРНК. Такой подход даёт возможность выявить именно те днРНК, паттерны экспрессии которых сходны с функционально значимыми мРНК, и, как следствие, определить функциональную значимость выявленных днРНК в регуляции клеточных процессов в исследуемых тканях.

Для оценки уровня экспрессии выявленных днРНК используется программа Kallisto v.0.46.2 (Bray, 2016). Рассматриваются два набора последовательностей: поливидовые и внутривидовые днРНК. Для расчета экспрессии пользователю нужно использовать библиотеки, на основе которых были собраны транскрипты. Расчет производится по следующей схеме. Если библиотеки принадлежат одной ткани, они объединяются и для них рассчитывается медиана экспрессии. Далее данные по всем тканям объединяют в одну таблицу. В сформированной таблице транскриптам присваиваются индексы, отражающие их принадлежность к одной из групп (поливидовые или внутривидовые днРНК). На основе полученной таблицы проводится анализ уровней экспрессии. Для каждой ткани с помощью критерия Стьюдента сравнивали распределения уровней экспрессии (медианных значений TPM) между группами транскриптов: поливидовые днРНК и внутривидовые днРНК.

Для оценки того, различаются ли общие уровни экспрессии между поливидовыми и внутривидовыми днРНК, для каждой ткани отдельно проводили попарное сравнение распределений значений TPM с помощью критерия Стьюдента и рассчитывали показатель статистической значимости различий - *p-value*, поскольку число сравниваемых тканей невелико.

Оценка тканеспецифичности производится отдельно для каждой из трёх групп транскриптов: поливидовых днРНК, внутривидовых днРНК и мРНК. Для

этого готовится база данных, в которой представлены проанализированные библиотеки и соответствующие им ткани (библиотеки с пометкой ‘unknown’ исключаются). После этого количество транскриптов каждой группы РНК, экспрессирующихся в данной ткани, нормируется на общее количество транскриптов данной группы РНК и на количество библиотек, соответствующих данной ткани. Предложенная нормировка позволяет учесть неравномерность числа генов, представленных в трех группах РНК, и неравномерность количества библиотек, соответствующих разным типам тканей. Таким образом предложенная метрика, является аналогом второй группы метрик оценивающих тканеспецифичность (см. пункт 1.4), то есть детальный анализ и расчет степени специфичности гена индивидуально для каждой конкретной ткани.

2.1.5 Платформа для реализации конвейера

Конвейер ICAAnnoLncRNA создан с применением программного менеджера Snakemake v6.0.0 (Köster, 2021). Система управления рабочим процессом Snakemake - это инструмент для создания воспроизводимых и масштабируемых конвейеров для анализа данных. Рабочие процессы описываются с помощью удобочитаемого языка на основе Python. Их можно легко масштабировать для серверов, кластеров, сетевых и облачных сред без необходимости изменять определение рабочего процесса. Благодаря совместимости Snakemake с системой управления Conda можно без труда устанавливать новые программы, необходимые для конвейера. Ссылка на конвейер: <https://github.com/artempronzin95/ICAnnoLncRNA-identification-classification-and-annotation-of-LncRNA>

2.1.6 Описание данных используемых для проверки работы биоинформатического конвейера ICAAnnoLncRNA

В качестве данных для проверки разработанного биоинформатического конвейера ICAAnnoLncRNA использовали 15 транскриптомных библиотек кукурузы (ссылки на используемые библиотеки приведены в таблице ПЗ). Мы применили два метода сборки транскриптома, чтобы продемонстрировать, что ICAAnnoLncRNA позволяет работать с любыми типами сборок.

Первый метод - сборка *de novo*, включает следующие шаги (Мухин, 2020): 1) извлечение прочтений из SRA-файла с помощью пакета SRA Toolkit (Staff, 2011); 2) фильтрация прочтений при помощи программы fastp v0.22.0 (Chen, 2018); 3) сборка транскриптов с использованием программы Trinity v2.6.6 (Haas, 2013); 4) оценка уровня экспрессии транскриптов с помощью программы Kallisto v.0.46.2 (Bray, 2016). Для оценки уровня экспрессии транскриптов использовали метод нормировки Transcripts Per Million (TPM).

Второй метод - сборка на основе референсного генома. Фильтрация и картирование прочтений на референсный геном кукурузы (версия AGPv4 базы данных Ensembl Plants (Jiao, 2017)) проводилась с помощью программы Hisat2 v.2.1 (Kim, 2015), сборка транскриптома - с помощью программы StringTie v.1.3 (Pertea, 2015). Данные этапы были реализованы с помощью биоинформатического конвейера LncPipe (Zhao, 2018). Аннотация референсного генома Ensembl Plants версии 40 также загружена с этого сайта.

Дополнительно мы провели проверку работы конвейера на *A. thaliana*. В качестве проверочных данных мы использовали 840 транскриптомных библиотек для *A. thaliana*, загруженных из открытых источников (Yuan et al., 2024). Не все библиотеки, полученные нами, содержали полную аннотацию экспериментов по анализу транскриптома (тип ткани, генотип и т.п.). Поэтому для анализа экспрессии днРНК в различных тканях и органах растения мы отобрали по 10 библиотек (всего 50), биологические образцы которых были получены из цветка, листа, корня, пыльника и побега. Для этих 50 библиотек мы оценили уровни экспрессии транскриптов, в том числе и днРНК.

Для данной проверки в качестве референсного генома использовался геном *A.thaliana*, представленный в базе данных TAIR (Garcia-Hernandez, 2002). Использовалась версия сборки генома от 11.07.2019 - TAIR10.1.

2.2 Разработка конвейера РТА для анализа пан-транскриптома

На сегодняшний день существует достаточное количество программ, направленных на сборку пан-геномов и пан-транскриптомов. Однако большинство из них создано для узких задач. Например, PGAP (Zhao et al., 2012) осуществляет масштабный поиск генов, проводит функциональную аннотацию,

обогащение кластеров ортологичных генов терминами генной онтологии, но не производит саму сборку пан-генома. Пакет программ PpsPCP (Tahir Ul Qamar et al., 2019) разработан для идентификации вариаций наличия/отсутствия генов (PAV) в пан-геномах, но не позволяет проводить построение ортогрупп последовательностей генов. Также ни одна программа не позволяет предсказывать и анализировать днРНК в масштабах пан-транскриптома.

В связи с этим нами было принято решение разработать собственный конвейер РТА (Pan-Transcriptome Assembler) для сборки и анализа пан-транскриптомов (рисунок 2.3).

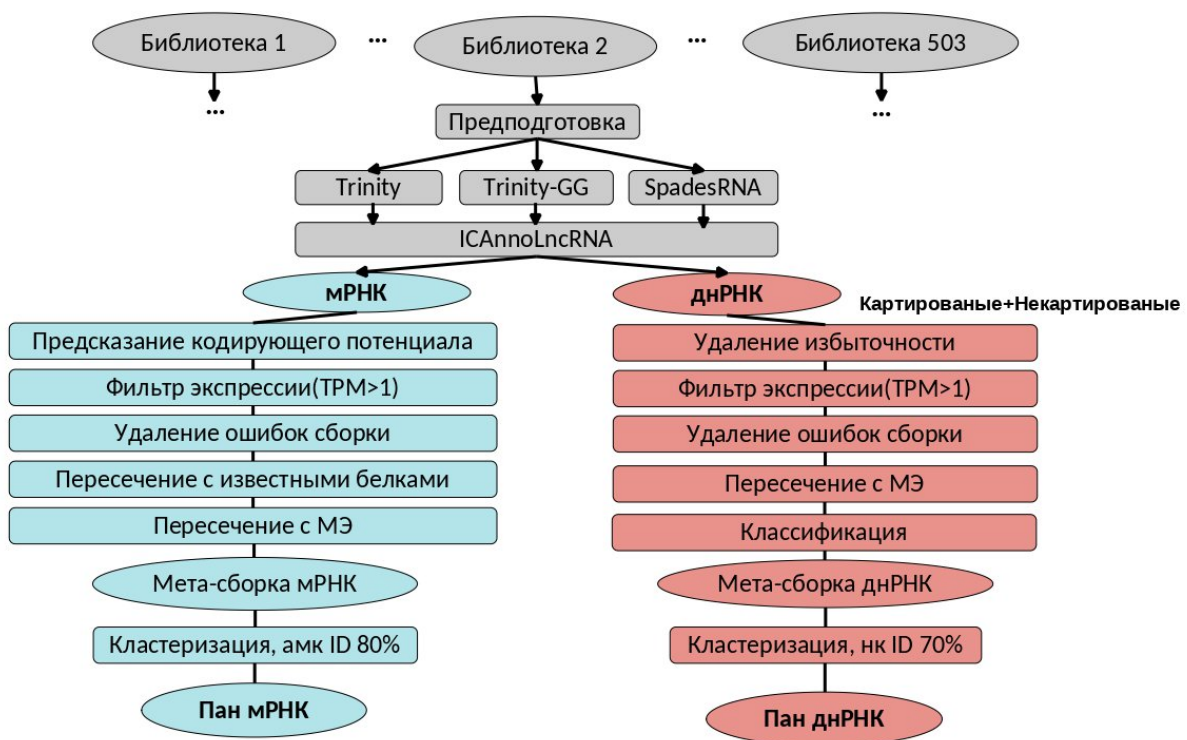


Рисунок 2.3. Схема конвейера анализа пан-транскриптома. Черные прямоугольники обозначают этапы анализа, общие как для мРНК, так и для днРНК. Красными прямоугольниками показаны этапы обработки, специфические для днРНК, синими - для мРНК. ID - процент идентичности при кластеризации.

Конвейер включает следующие этапы обработки: сборка транскриптома для отдельных библиотек тремя методами сборки, классификация транскриптов трех сборок отдельных линий на кодирующие и некодирующие, объединение результатов в мета-сборки индивидуальных линий, сборка на основе этих данных

пан-транскриптомов для мРНК и днРНК, анализ структуры, вариаций и скорости нуклеотидных замен пан-транскриптома. Детальный анализ каждого этапа представлен ниже.

2.2.1 Сборка транскриптомов

Для контроля качества исследуемых библиотек применяется программа FastQC v0.11.8 (Chen, 2018). Для фильтрации прочтений используется fastp v0.22.0 (Chen, 2018) со следующими параметрами: среднее качество секвенирования (параметр Phred) - 20, минимальная длина прочтения - 50, длина перекрытия - 50.

Сборка транскриптома производится программами Trinity, Trinity Genome-guided (TrinityGG) и SPAdes независимо для каждой библиотеки. Данные программы были выбраны из-за высоких показателей качества и точности сборки транскриптомов *de novo* (Hölzer, 2019; Shmakov, 2021; Jain, 2013).

Сборка Trinity v2.9.1 (Grabherr, 2011) производится с параметрами по умолчанию. Для сборки с помощью SPAdes v3.13.0 (Prjibelski, 2020) используется режим SPAdes-rna, также с параметрами по умолчанию.

Для сборки транскриптома на основе референсного генома прочтения картируются на референсный геном выбранного организма с помощью программы Hisat2 v2.2.1 (Kim, 2019). С помощью samtools v1.6 (Danecek, 2021) производится предобработка и сортировка полученных файлов картирования. Предобработанные результаты картирования подаются на вход программе Trinity в режиме Genome-guided (Venturini, 2018); максимальная длина интрона ограничивалась значением 500 000 п.н.

Далее для каждой из трех полученных сборок отдельных библиотек производится разделение на кодирующую и не кодирующую часть с помощью описанного выше конвейера ICApnoLncRNA (Pronozin, 2023). В результате работы конвейера транскрипты для независимых сборок разбиваются на кодирующие и не кодирующие последовательности. Дальнейший их анализ происходит параллельно.

2.2.2 Сборка пан-транскриптома не кодирующих последовательностей

Идентификация днРНК картированных на референсный геном

Все описанные этапы фильтрации днРНК, представленные на рисунке 2.3, производятся конвейером ICAAnnoLncRNA, детали которых были описаны в разделе 2.1.

Идентификация днРНК некартированных на референсный геном

Существенная доля транскриптов днРНК не была картирована на референсный геном. Для того чтобы данные транскрипты также использовать в исследовании, были предприняты следующие шаги. Сформирована выборка известных белок кодирующих генов и некодирующих РНК (мРНК, микроРНК, тРНК, межгенные РНК) из БД Ensembl Plants (Bolser et al., 2016). Дополнительно использованы библиотеки мобильных элементов генома кукурузы, полученные в работе (Pronozin, 2023). Транскрипты, не имеющие сходства с фрагментами референсного генома, сравнивались с указанными последовательностями с помощью программы Blastn v.2.5.0 (Boratyn et al., 2013). Считалось, что транскрипт содержит фрагмент мобильного элемента, если было идентифицировано 80% идентичности с известными мобильными элементами кукурузы. Для выявления гомологии с белок кодирующими генами использовались параметры «qcovs» (показывает процент покрытия query-последовательности относительно target) и «sstrand» (показывает направление query-последовательности относительно target); таким образом, транскрипт считался белок кодирующим при условии qcovs=100%,sstrand=plus (то есть направлен в плюс-цепи). Для выявления гомологии с некодирующими РНК использовался параметр qcovs=100%. Последовательности, имеющие гомологию с белок кодирующими генами, мобильными элементами и некодирующими РНК, удалялись из исследуемого набора.

Полученный набор некартированных днРНК мы объединили с набором картированных днРНК, отдельно для каждой линии кукурузы.

Сборка пан-транскриптома днРНК

Полученные наборы транскриптов днРНК из отдельных сборок объединяются в мета-сборки. Мета-сборка позволяет компенсировать недостатки отдельных сборщиков, улучшая общее качество сборки (Cerveau, 2016; Evangelistella, 2017; Wences, 2015; Mirebrahim, 2015; Shmakov, 2021). Мета-

сборки для библиотек формируются с помощью программы `trformat.pl` из пакета *Evidential Gene* (Gilbert, 2019). Полученные мета-сборки также объединяются в одну общую сборку. Последовательности в этой сборке кластеризуются с помощью *VSEARCH v2.15.2* (Rognes, 2016) для идентификации репрезентативных транскриптов (центроидов). При кластеризации днРНК учитывалось, что даже для одной и той же пары мРНК идентичность нуклеотидных и аминокислотных последовательностей различается (Kenneth, 1993). Следовательно, применение одного и того же порога сходства (80%) для кластеризации как днРНК, так и аминокислотных последовательностей привело бы к различиям в структурах двух пан-транскриптомов из-за неравенства скоростей дивергенции. Поэтому была проведена эмпирическая процедура выбора порога кластеризации для днРНК, приблизительно соответствующего 80% сходства аминокислотных последовательностей. Случайным образом было выбрано 100 мРНК из разных кластеров. Последовательности из одних кластеров выравнивались с помощью *MACSE v2* (Ranwez et al., 2018) с сохранением соответствия кодонов. Нуклеотидное различие оценивалось для пар последовательностей с аминокислотным различием, близким к 0, 10, 20, 30 и 40%. Результирующие распределения представлены в виде столбчатой диаграммы на рисунке 2.4.

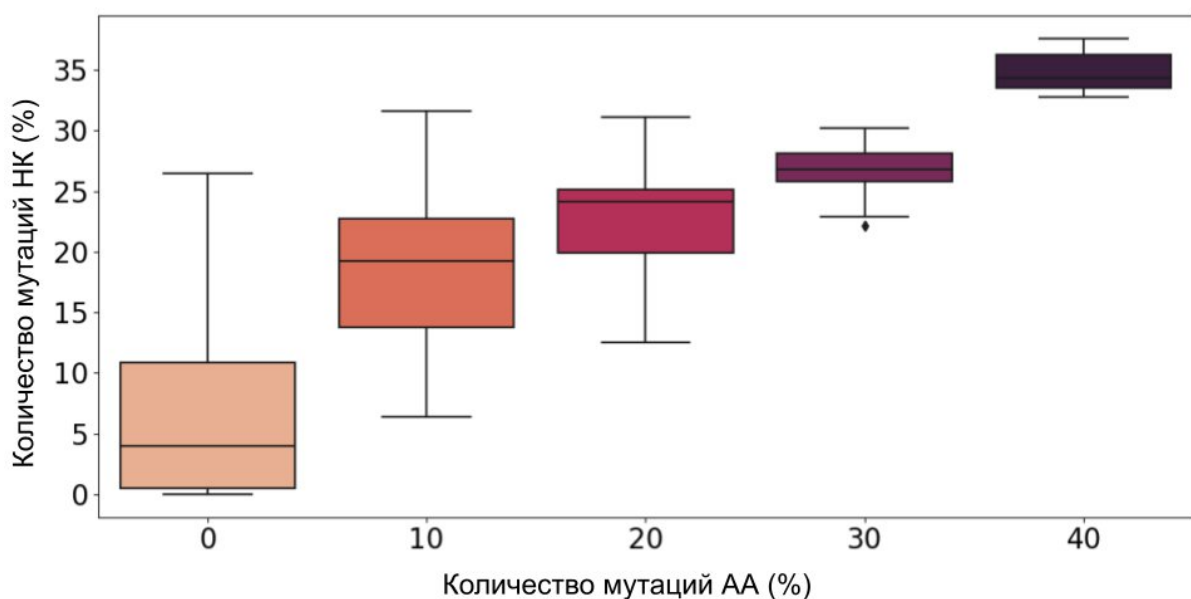


Рисунок 2.4. Распределение количества аминокислотных мутаций относительно нуклеотидных мутаций в последовательностях транскриптов из исследуемых кластеров. По оси X отложена доля мутаций при кластеризации аминокислотных последовательностей, по оси Y доля мутаций при кластеризации нуклеотидных последовательностей.

На рисунке показано, что при 20 % различии белковых последовательностей третий квартиль распределения нуклеотидных различий близок к 30% (медиана 24.5%). Поэтому для кластеризации днРНК был использован порог идентичности 70%. Кластеры транскриптов днРНК исключались из анализа, если уровень экспрессии их центроидов был ниже 1 TPM. Из центроидов, прошедших фильтрацию собирается пан-транскриптом днРНК.

2.2.3 Сборка пан-транскриптома кодирующих последовательностей

Кодирующие последовательности трех независимых сборок обрабатываются программой `tr2aacds.pl` из пакета Evidential Gene (Gilbert, 2019) для устранения избыточности и контаминации. Для отобранных транскриптов с помощью Kallisto v.0.46.2 (Bray, 2016) оценивается уровень экспрессии; транскрипты с низким уровнем экспрессии ($TPM < 1$) исключаются из дальнейшего анализа. Для отобранных транскриптов с помощью программы Evidential Gene формируются соответствующие им аминокислотные последовательности.

Для исключения контаминации аминокислотные последовательности сравниваются с последовательностями NCBI Univec (Federhen, 2012) с помощью программы Blastx v.2.5.0 (Boratyn, 2013). Значимые совпадения исключаются из анализа. Затем проводится сравнение последовательностей с базой данных PLAZA (Van Bel, 2018) (учитывались последовательности однодольных растений *Z. mays* B73, *Saccharum spontaneum*, *S. bicolor*, *O. sativa* ssp. *indica*, *O. sativa* Kitaake, *O. sativa* ssp. *japonica*, *B. distachyon*, *H. vulgare*, *T. dicoccoides*, *T. turgidum*, *T. aestivum*). Последовательности библиотек исследуемого организма

исключаются из дальнейшего анализа в случае, если уровень сходства с последовательностями из PLAZA ниже 90%.

На основе отфильтрованных последовательностей мРНК трех сборок формируется мета-сборка.

Последовательности мета-сборок отдельных библиотек объединяются в пан-транскриптом кодирующей части. Для определения состава генов пан-транскриптома их последовательности объединяются в кластеры на основе кластеризации транслированных аминокислотных последовательностей с использованием программы usearch v.11 (Edgar, 2010) при пороге идентичности 80%. Из полученных центроидов формируется пан-транскриптом белок кодирующих генов.

2.2.4 Анализ структуры пан-транскриптома

На основе наличия или отсутствия транскриптов от разных линий/образцов, принадлежащих общим кластерам, производится анализ зависимости размера пан-транскриптома и его ядра от количества организмов в пан-транскриптоме с помощью алгоритма PanGP (Zhao, 2014). Результаты выводятся для двух типов пан-транскриптома в виде отдельных графиков.

Для классификации транскриптов по представленности в пан-транскриптоме используются следующие пороги: к группе ядра относятся транскрипты, присутствующие в 100% транскриптомов; к группе оболочки - представленные в 99-10% транскриптомов; к группе облака - в менее чем 10% транскриптомов.

Для определения типа пан-транскриптома на основе данных о росте числа новых генов в зависимости от числа представителей вида использовались параметры степенного закона (Tettelin, 2008): α_{pan} - показатель степени, $tg(\theta_{pan})$ - асимптотический прирост числа новых генов в пан-геноме/пан-транскриптоме при добавлении нового представителя. Первый параметр позволяет определить тип пан-транскриптома: если $\alpha_{pan} < 1$, то он считается открытым; если $\alpha_{pan} > 1$, то считается закрытым.

2.2.5 Анализ нуклеотидных вариаций последовательностей пан-транскриптомов

Чтобы оценить нуклеотидные различия последовательностей мРНК и днРНК в исследуемых организмах, используется оценка параметров «нуклеотидное разнообразие», π (Nucleotide diversity), и показателя теста Таджимы D (Tajima's D), а также строится их распределение для мРНК и днРНК (Wright, 2005; Carlson, 2005; Akhunov, 2010). Параметр π является мерой генетического разнообразия в популяции. Он показывает уровень вариаций нуклеотидов между представителями популяции. Чем выше значение π , тем больше разнообразие нуклеотидов и, соответственно, больше однонуклеотидных вариаций. Параметр D оценивает разницу между средним числом парных нуклеотидных различий и числом сегрегирующих позиций и позволяет определить, насколько распределение аллелей отличается от того, что ожидалось при нейтральной эволюции. При $D \approx 0$ замены соответствуют модели нейтральной эволюции. При $D < 0$ наблюдается больше редких аллелей, чем в случае нейтральной эволюции, что может указывать на положительный отбор или эффект недавнего расширения популяции. При $D > 0$ наблюдается меньше редких аллелей, чем в случае нейтральной эволюции, что может указывать на отрицательный (очищающий) отбор.

Параметры π и D оцениваются с помощью программы vcftools v.0.1.16 (Danecek, 2011), которая в качестве входных данных использует файлы геномных вариаций в формате VCF. Для их получения, в свою очередь, используется картирование прочтений из индивидуальных библиотек на последовательности центроидов кластеров, полученных при анализе пан-транскриптомов. Центроиды объединяются в две псевдо-хромосомы. Первая включает только центроиды мРНК, вторая - днРНК. В псевдо-хромосомах отдельные транскрипты объединяются в единую нуклеотидную последовательность, между последовательностями транскриптов добавляется вставка из 300 символов N, чтобы исключить картирование прочтений длиной 100 п.н. на оба центроида одновременно. Прочтения библиотек картируются на псевдо-хромосомы с помощью программы bwa mem v.0.7.17 (Li, 2013). Поиск полиморфизмов (ОНП,

вставок и делеций (индел)) производится с помощью комбинации программ samtools mpileup и bcftools call (Danecek, 2021). Для полученного набора полиморфизмов применяется фильтрация на основе $MAPQ > 30$ и missing value = 0.25, $MAF = 0.05$, $HWE = 0.01$ и $LD = 0.2$. Позиции транскриптов с заменами, прошедшие данный фильтр, используются при расчете параметров π и D . Для оценки статистической значимости различий между группами белок кодирующих генов и днРНК используется тест Манна - Уитни.

2.2.6 Анализ синонимических и несинонимических замен

Чтобы оценить режим эволюции последовательностей мРНК и днРНК в популяциях исследуемых организмов, используется оценка скоростей несинонимических (K_a) и синонимических (K_s) замен для мРНК и нуклеотидных замен (K_n) для днРНК. Эти параметры оцениваются с помощью программы KaKs_calculator v3 (Zhang, 2022), которая на вход использует парные выравнивания нуклеотидных последовательностей с референсными. Поэтому для анализа замен в кодирующих последовательностях используются последовательности CDS референсного генома, которые могут быть загружены из любой доступной базы данных.

Для анализа днРНК используются последовательности PLncDB v.2.0 (Jin, 2021). Данная база данных объединяет сразу несколько БД днРНК, включая EVlncRNAs (Zhou, 2018) (БД с экспериментально подтвержденными днРНК), а также предоставляет уровень доверия находящимся в ней данным. Так, днРНК делятся по уровню доверия к ним от низкого к высокому. В результате отбираются только днРНК из БД EVlncRNAs и днРНК с высоким уровнем доверия. Далее с помощью программы GMAP (Wu, 2005) производится выравнивание транскриптов для каждой из сборок, принадлежащих образцам, анализируемым пользователем на референсные последовательности CDS и днРНК.

Анализ производится для последовательностей отдельных библиотек. Для мРНК оценивается распределение скоростей K_a , K_s и их отношения K_a/K_s для каждого транскрипта. Для днРНК оценивается распределение параметра K_n и его отношения к среднему значению K_s для всех мРНК транскриптома.

Для оценки режима замен в последовательностях мРНК и днРНК гистограммы K_a , K_s , K_n аппроксимируются гамма-распределением, параметрами которого являются α и β (Yang, 1996). Первый параметр оценивает характер замен: если он меньше 1, то последовательности консервативны и содержат лишь незначительное число сайтов, в которых скорость замен высокая. Если $\alpha > 1$, это означает, что в последовательности представлены сайты с широким разнообразием по скоростям замен (Yang, 1996). Сравнение распределений K_a , K_s , K_n проводилось для мРНК и днРНК, а также для последовательностей, разделенных по группам пан-транскриптома: ядро, оболочка, облако.

2.2.7 Транскриптомные данные инбредных линий кукурузы

В качестве исследуемых данных мы использовали 503 транскриптомные библиотеки инбредных линий кукурузы из работы (Hirsch et al., 2014). Авторы поставили перед собой цель расширить пределы одного референсного генома и получить полный набор генов кукурузы.

Панель линий включала 465 разнообразных американских инбредных линий кукурузы из базы данных Wisconsin Diversity Set (Hansey et al., 2010) и 38 инбредных линий, полученных авторами работы (Hirsch et al., 2014), рисунок 2.5.

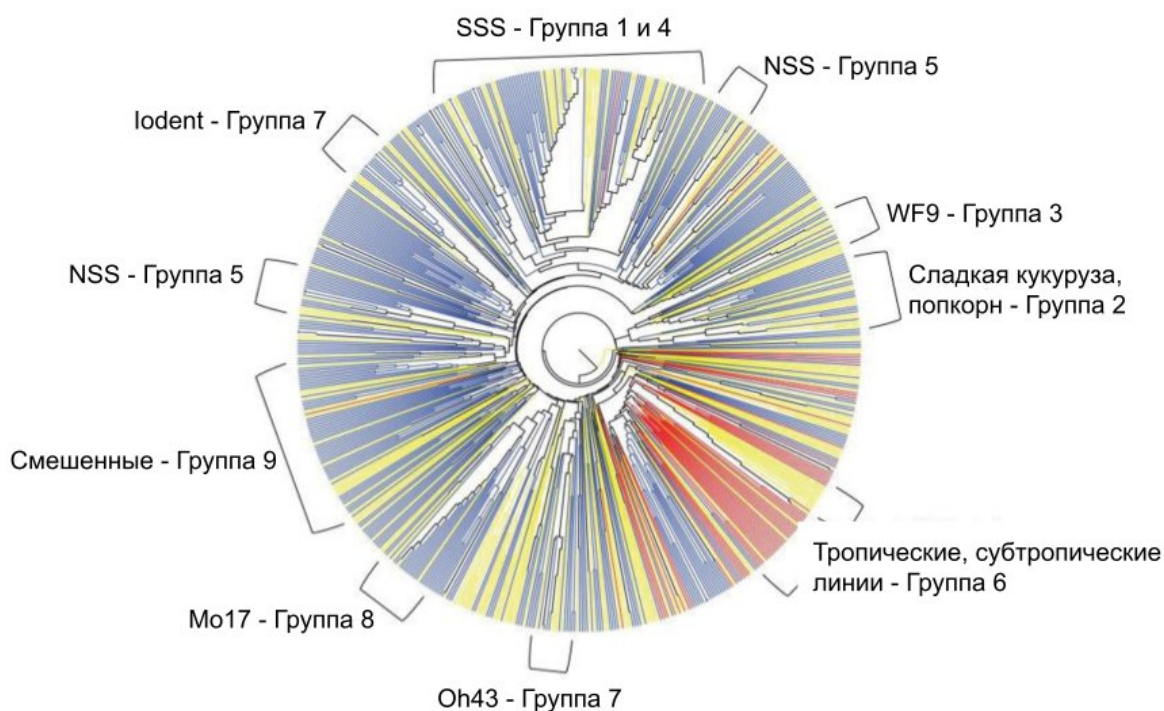


Рисунок 2.5. Панель видов включала 465 разнообразных американских линий

кукурузы из Wisconsin Diversity Set и 38 инбредных линий, полученных авторами работы.

Все растения выращивались в контролируемых тепличных условиях (27 °C/24 °C день/ночь, фотопериод 16/8 ч). Сбор проростков производили на стадии V1 (молодые листья рассады) (Abendroth et al., 2011); брали целый проросток, включая корни. Для каждой линии объединяли по три растения на генотип для максимизации разнообразия транскриптов. Тотальную РНК выделяли фенол-хлороформным методом с использованием реагента Trizol с последующей очисткой от примесей и остаточного фенола, а также дополнительно концентрировали, используя набор Qiagen RNeasy MinElute Cleanup Kit. Затем дважды провели поли(А)-селекцию с использованием Dynabeads mRNA Isolation Kit (Invitrogen) для удаления рибосомальной фракции. Секвенирование выполняли на Illumina HiSeq (100 п.н., парные концы).

Работа была разделена на два основных блока: поиск новых генов и оценка размеров собранного пан-транскриптома. Под поиском новых генов (рисунок 2.6) авторы подразумевают сбор и аннотацию генов, не представленных в референсном геноме кукурузы B73. Для этого все 503 библиотеки были картированы на референсный геном; прочтения, не имеющие картирования, были использованы для сборки транскриптома *de novo*. Сборка производилась программами Velvet (Zerbino и Birney, 2008) и Oases (Schulz et al., 2012). В результате сборки было получено 102 017 транскриптов, которые с помощью алгоритмов Oases были сгруппированы в 31 398 локусов. Для дальнейшего анализа внутри каждого локуса выбирали один транскрипт «representative transcript assembly» (RTA). В качестве RTA авторы брали самый длинный транскрипт. Для полученного набора RTA проводилась фильтрация. Удаляли транскрипты, имеющие картирование с референсным геномом, и возможные загрязнения вирусной РНК. Любой RTA, который картируется на геном с покрытием не менее 85% длины и идентичностью не менее 85%, считали уже существующим в референсном геноме. Загрязнения вирусной РНК удалялись путём сравнения с базой данных Univec (Federhen, 2012). Также для оставшихся

RTA проводился поиск гомологов среди известных белков кукурузы и злаковых (*O. sativa* и *S. bicolor*).

Поиск новых генов

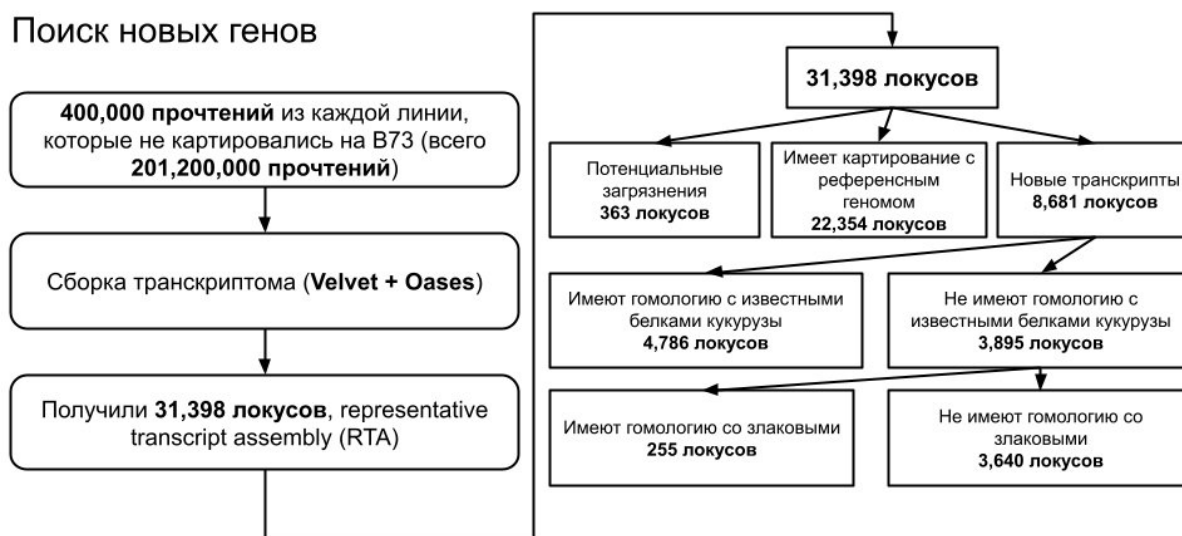


Рисунок 2.6. Схема выявления новых белок кодирующих генов и дальнейшая их аннотация.

Оценка размеров собранного пан-транскриптома (рисунок 2.7). На данном этапе производится оценка типа пан-транскриптома (открытый или закрытый). Для этого авторы отобрали 366 библиотек, имеющих более трёх миллионов некартированных прочтений. Для каждой из 366 линий был собран *de novo* транскриптом с помощью комбинации программ Velvet и Oases. Затем все сборки были объединены и картированы на референсный геном с целью удалить известные гены кукурузы. В результате авторы получили 747 977 RTA-транскриптов. Полученные RTA были распределены по ортологическим группам с помощью программы OrthoMCL (Li, 2003).

Оценка размера пан-генома

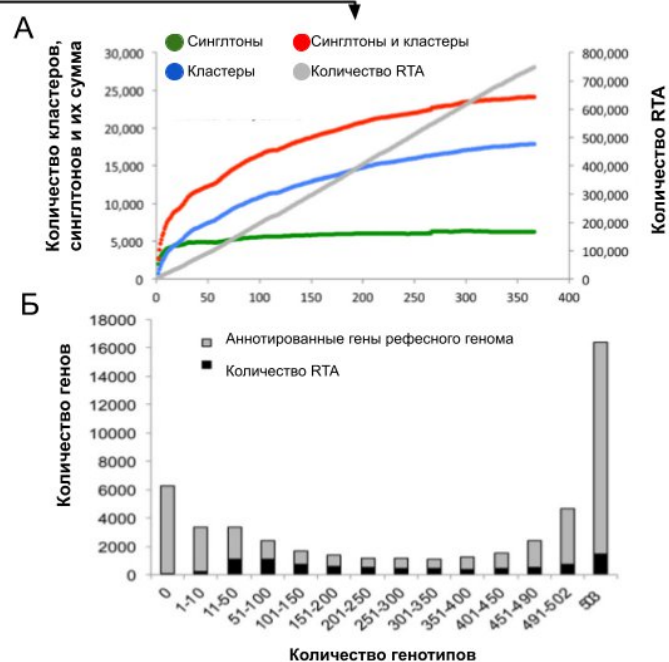
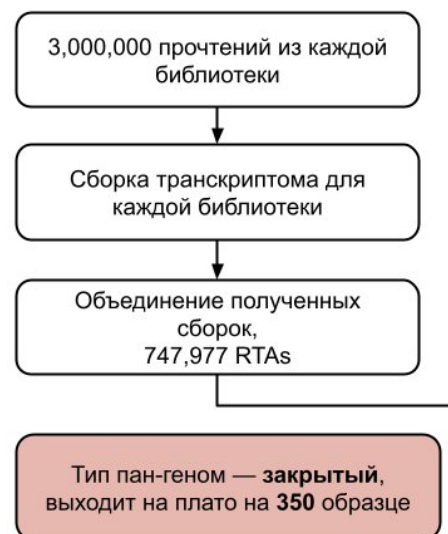


Рисунок 2.7. Схема оценки структуры пан-транскриптома 366 линий кукурузы. А) Оценка типа пан-транскриптома на основе единичных кластеров (синглтонов), полноценных кластеров и их суммы. Б) Оценка распределения экспрессии генов по 503 инбредным линиям кукурузы. Серым показаны аннотированные гены кукурузы из референсного генома, черным RTA транскрипты.

Оценка открытости/закрытости пан-транскриптома производилась на основе полученных ортологических групп (рисунок 2.7А). На первом шаге случайным образом выбирались две линии из 366, далее объединяли их RTA, кластеризовали и подсчитывали, сколько ортогрупп сформировалось. На втором шаге добавляли случайную третью линию и проводили тот же процесс. Добавление продолжалось до последней, 366-й линии. Данный процесс повторяли несколько раз. В результате было показано, что общее количество ортологических групп выходит на плато на 350-й линии.

Также авторы провели анализ распределения экспрессии аннотированных генов из референсного генома кукурузы и выявленных ими новых RTA-транскриптов. Цель состояла в том, чтобы показать, сколько транскриптов и генов присутствуют во всех 503 линиях, а сколько — только в их части. В результате 14 968 аннотированных генов В73 и 1425 RTA экспрессируются во

всех 503 линиях и, следовательно, принадлежат ядру. Тогда как 18 327 генов B73 и 7183 RTA относятся к облаку и оболочке (авторы не разделяли их по отдельности).

В результате данная работа показала, что пан-транскриптом значительно расширяет набор генов по сравнению с одним референсным геномом. Авторы продемонстрировали, что на ядро приходится меньшая доля генов, чем на переменную часть. Это свидетельствует об отсутствии в референсном геноме генов, присущих некоторым уникальным линиям кукурузы.

В качестве референсной использовалась последовательность генома кукурузы v. 40 (AGPv4) (Jiao, 2017), аннотация была взята из БД Ensembl Plants v.50 (Howe et al., 2020).

2.2.8 Проверка качества сборки транскриптомов

Данная проверка осуществляется с целью оценить, насколько качественно выполняется сборка транскриптомов для каждой линии кукурузы с помощью разработанного нами конвейера RTA. Оценка основывалась на анализе характеристик картирования и показателей соответствия между собранными транскриптами и аннотированными последовательностями референсного генома.

В качестве тестового образца для оценки качества сборки транскриптома мы использовали мета-сборку транскриптов референсной линии кукурузы B73. Полученный набор транскриптов сравнивали с аннотированными генами референсного генома B73 с помощью rnaQUAST (Bushmanova et al., 2016). Такой подход был выбран, поскольку в исходной работе (Hirsch et al., 2014) сборка индивидуальных геномов для каждой линии кукурузы не производилась, и референсный геном B73 является единственным доступным эталоном для контроля качества. Для оценки качества сборки были выбраны следующие параметры:

- Картированные (Aligned) - количество транскриптов с хотя бы одним значимым выравниванием на геном, отражающее долю успешно картированных последовательностей.

- Некартированные (Unaligned) - количество транскриптов без выравниваний на геном.
- Ошибки сборки (Misassemblies) - число транскриптов, которые считаются ошибками сборки на основе двух методов картирования. С помощью данной метрики можно выявить транскрипты, полученные ошибочным соединением фрагментов из разных генов.
- Покрытие аннотированных генов (Database coverage) - процент аннотированных генов референсного генома, для которых в сборке найден хотя бы один значимо выровненный транскрипт.
- Средняя доля картированных фрагментов (Avg. aligned fraction) - средняя доля оснований в транскриптах, которые выравниваются на геном.

Также подобную оценку мы провели для днРНК, полученных для той же линии кукурузы В73.

ГЛАВА 3 Применение конвейера IAnnoLncRNA для анализа транскриптомов кукурузы и арабидопсиса

3.1 Результаты анализа транскриптомов кукурузы

3.1.1 Обучение и валидация модели предсказания днРНК

В данной работе для обучения модели предсказания днРНК использовался набор из 2535 последовательностей известных днРНК и 33 725 мРНК кукурузы, представленных в аннотации Ensembl v. 40 (Jiao, 2017). Как было указано в пункте 2.1.1 материалов и методов, разработанный конвейер позволяет проводить предсказание в двух режимах. Предварительный анализ показал, что использование параметров модели, включающей особенности вторичной структуры последовательностей, требует в 16 раз больше затрат времени (3 часа 58 минут против 15 минут соответственно) в сравнении с моделью, не использующей вторичную структуру. Это обусловлено использованием программы предсказания вторичной структуры ViennaRNA (Lorenz et al., 2011). При этом точность распознавания по нашим данным увеличивалась незначительно: $F1 = 0.951$ против $F1 = 0.960$ (примерно на 1%). Поэтому в настоящей работе мы не использовали параметры вторичной структуры для предсказания кандидатов в днРНК. Однако для других организмов точность может быть выше, поэтому мы добавили возможность включения/выключения данной опции.

Результаты оценки обучения модели для кукурузы представлены в таблице 3.1. Значение метрики $F1 = 0.951$ указывает на сбалансированность точности и полноты модели. Значение метрики $MCC = 0.86$, менее чувствительной к дисбалансу классов, отражает высокую степень соответствия предсказаний модели реальным классам.

Таблица 3.1: Оценка точности модели предсказания днРНК для последовательностей кукурузы

Метрика	Кратность валидации	Результаты, Среднее $\pm$ std
$F1$	5x	0.9512 ± 0.0029
	10x	0.9512 ± 0.0041

<i>MCC</i>	5x	0.8620 ± 0.0078
	10x	0.8621 ± 0.0112

Также стоит обратить внимание на низкие значения стандартного отклонения для обеих метрик. Это указывает на то, что между выборками незначительный разброс результатов и качество предсказания не зависит от случайного разбиения данных. Таким образом, модель обладает высокой обобщающей способностью и не переобучается на специфичных последовательностях отдельных выборок.

Результаты выбора протокола валидации (5-кратной и 10-кратной кросс-валидации), представлены на рисунке 3.1. Переход к 10-кратной валидации не привел к статистически значимому приросту средних значений метрик: показатели *F1* и *MCC* остались сопоставимыми (0.9512 против 0.9512 и 0.8620 против 0.8621 соответственно). При этом дисперсия результатов при 10-кратной валидации незначительно возросла.

В результате на основе анализа 5 реплик сформированных выборок для дальнейшего предсказания днРНК выбраны параметры модели LncFinder, которые обеспечивают наибольшую точность распознавания днРНК ($F1 = 0.955$ и $MCC = 0.870$). В данном случае это 1 итерация.

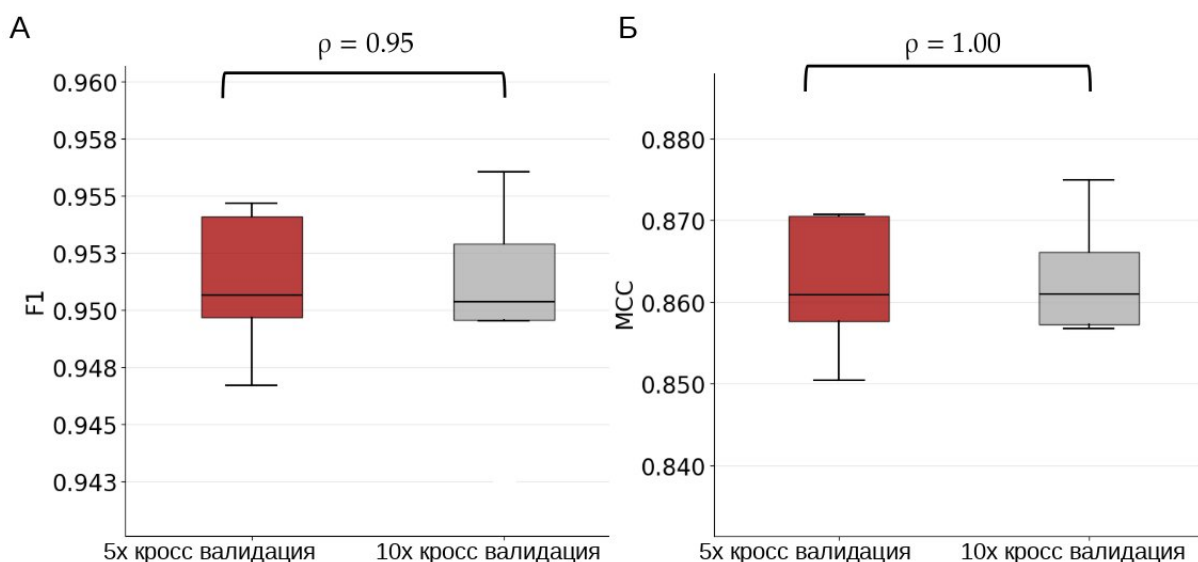


Рисунок 3.1. Сравнение 10 и 5-кратной кросс валидации для обучения модели предсказания днРНК. Ось Y - значение метрик *F1* и *MCC*. Точками показаны итерации кросс-валидации.

Поскольку статистических различий между средними показателями точности для 10-кратной и 5-кратной кросс-валидации не было выявлено, было решено использовать в дальнейшем вариант с 5 репликами.

Таким образом, полученные результаты подтверждают высокую точность и робастность модели предсказания днРНК, что позволяет применять ее для дальнейшего анализа этих последовательностей.

3.1.2 Идентификация днРНК

Вычислительный конвейер был использован для идентификации днРНК в 15 транскриптомных библиотеках кукурузы, собранных двумя способами: с помощью Trinity (*de novo*) и Hisat2/StringTie (картирование на референсный геном). Результаты каждого этапа работы конвейера ICAncRNA приведены в таблице 3.2. Из таблицы видна разница в количестве транскриптов между сборками, подаваемыми на вход конвейера (сборка Trinity имеет примерно в 6 раз больше транскриптов, чем сборка Hisat2/StringTie). Однако количество идентифицированных транскриптов днРНК различается еще сильнее: результаты для сборки Trinity примерно в 20 раз превышают результаты для сборки Hisat2/StringTie. Аналогичная ситуация наблюдается и для локусов.

Таблица 3.2: Идентификация днРНК с помощью разработанного биоинформатического конвейера ICAncRNA для 15 транскриптомных библиотек кукурузы, собранных двумя способами: Trinity и Hisat2/StringTie.

Этап конвейера	Trinity	Hisat2/StringTie
Входные данные	656 289	91 721
Фильтр по длине последовательности	656 289 (100%)	91 721 (100%)
LncFinder	450 504 (69%)	21 186 (23%)
Картирование	441 733 (67%)	18 428 (20%)
Фильтр ошибок/"шума"	101 207 (15%)	9837 (10%)
Фильтр пересечений с TE	40 203 (6%)	6128 (6%)
Кандидаты в днРНК	61 004 (9%)	3706 (4%)
Локусы кандидатов в днРНК	52 526 (8%)	3583 (4%)

В результате работы конвейера было выявлено 52 526 локусов днРНК для сборки Trinity и 3583 для сборки Hisat2/StringTie. Локусы днРНК занимают 19 525 698 п.н. (1% от размера референсного генома) для сборки Trinity и 1 799 826 п.н. (0.08 % от размера референсного генома) для сборки Hisat2/StringTie. Исходя из результатов, в среднем количество транскриптов, приходящихся на один локус, одинаково между сборками: 1.16 и 1.03 соответственно.

Распределение идентифицированных нами локусов днРНК для обеих сборок на хромосомах кукурузы показано на рисунке 3.2. Из рисунка 3.2 А, Б видно, что наибольшая плотность днРНК наблюдается для хромосом 1, 5, 8 для сборки Trinity и 1, 2, 8 для сборки Hisat2/StringTie. Наименьшая плотность наблюдается для хромосом 3, 4 в обеих сборках. Также из рисунка 3.2 В, Г видно, что распределение локусов днРНК по длине хромосом неравномерно. Частота встречаемости локусов днРНК выше на концевых участках хромосом и ниже на центромерных участках. Данная тенденция одинакова для обеих сборок.

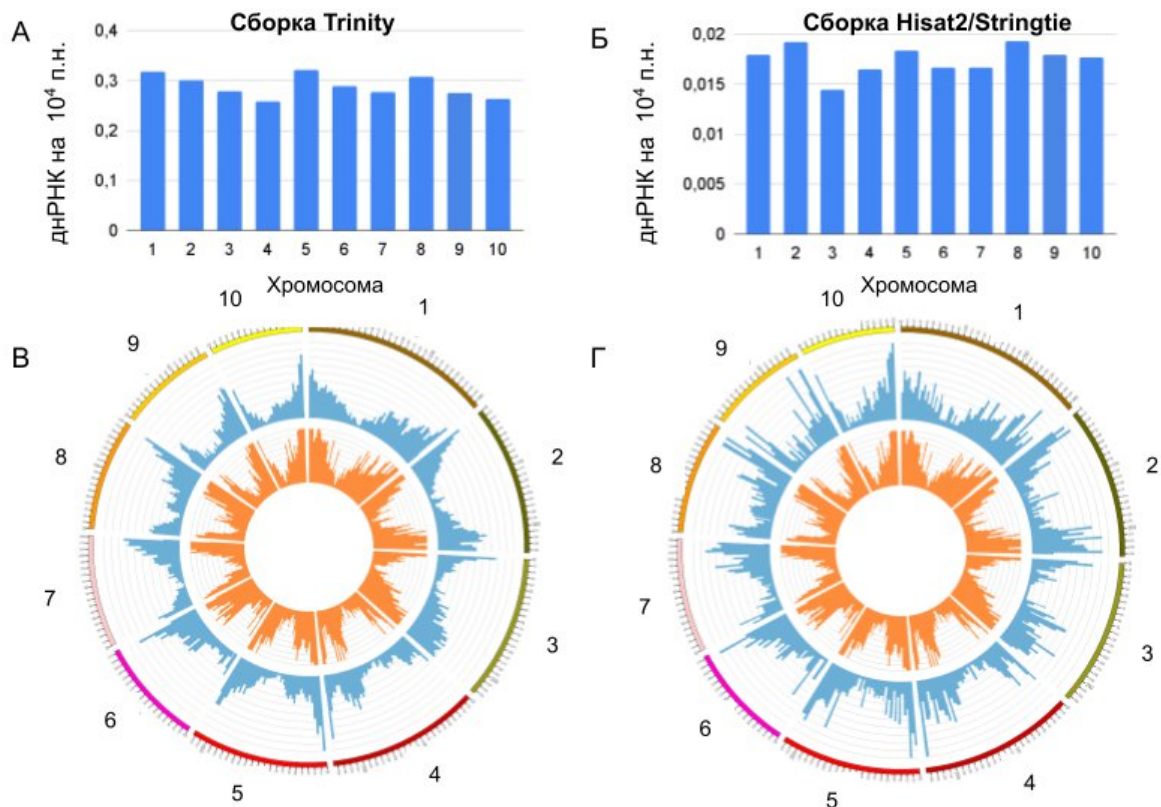


Рисунок 3.2. Распределение выявленных локусов днРНК по хромосомам кукурузы. А,Б) Плотность днРНК (ось Y) для каждой из хромосом (ось X). В,Г) Распределение локусов днРНК по длине хромосом в виде круговой диаграммы. Оранжевым цветом подсвечены - мРНК, синим - днРНК.

3.1.3 Классификация днРНК

Выявленные днРНК были классифицированы по их расположению в геноме относительно белок кодирующих генов (см. пункт 2.1.2) на три класса: межгенные, антисмысловые и интронные. Распределение выявленных днРНК по данным классам показано в таблице 3.3. Также в данную таблицу добавлены транскрипты, полностью перекрывающиеся с известными днРНК из базы данных Ensembl plants v.40, обозначаются как «=» (последняя строка таблицы).

Таблица 3.3: Распределение транскриптов, предсказанных как днРНК, на классы.

Классы	gffcompare	Сборка Trinity	Сборка Hisat2/StringTie

антисмысловые	x	34611 (56%)	1590 (42%)
интронные	i	17043 (27%)	180 (4%)
межгенные	u	9350 (15%)	1936 (52%)
известные днРНК	=	306	297

Соотношение антисмысловых, интронных и межгенных днРНК составляет 56:27:15 для сборки Trinity и 42:4:52 для сборки Hisat2/StringTie. Заметно, что второй метод сборки дает гораздо меньшую долю предсказанных интронных днРНК и большую долю межгенных днРНК. Интересно, что количество транскриптов, совпавших с известными днРНК, близко для обоих методов сборки (306 и 297).

3.1.4 Анализ структурных характеристик днРНК

Выравнивание транскриптов на референсный геном позволило идентифицировать экзон-интронную структуру выявленных локусов, кодирующих днРНК. Распределение локусов по количеству экзонов, длине экзонов и интронов представлено на рисунке 3.3.

Анализ показал, что более 80% днРНК содержат лишь один экзон для обеих выборок (рисунок 3.3 А,Б). Однако сборка Hisat2/StringTie демонстрирует значительную долю локусов, имеющих два и три экзона (более 10% для локусов днРНК с двумя экзонами). Распределение длин экзонов днРНК имеет два характерных региона. Небольшая доля экзонов имеет длину порядка 200 п.н., однако большинство днРНК имеют длину экзона в пределах от 300 до 1000 п.н. для обеихборок.

Распределение интронов по длинам имеет один характерный пик, расположенный на 100 п.н. для обеих выборок.

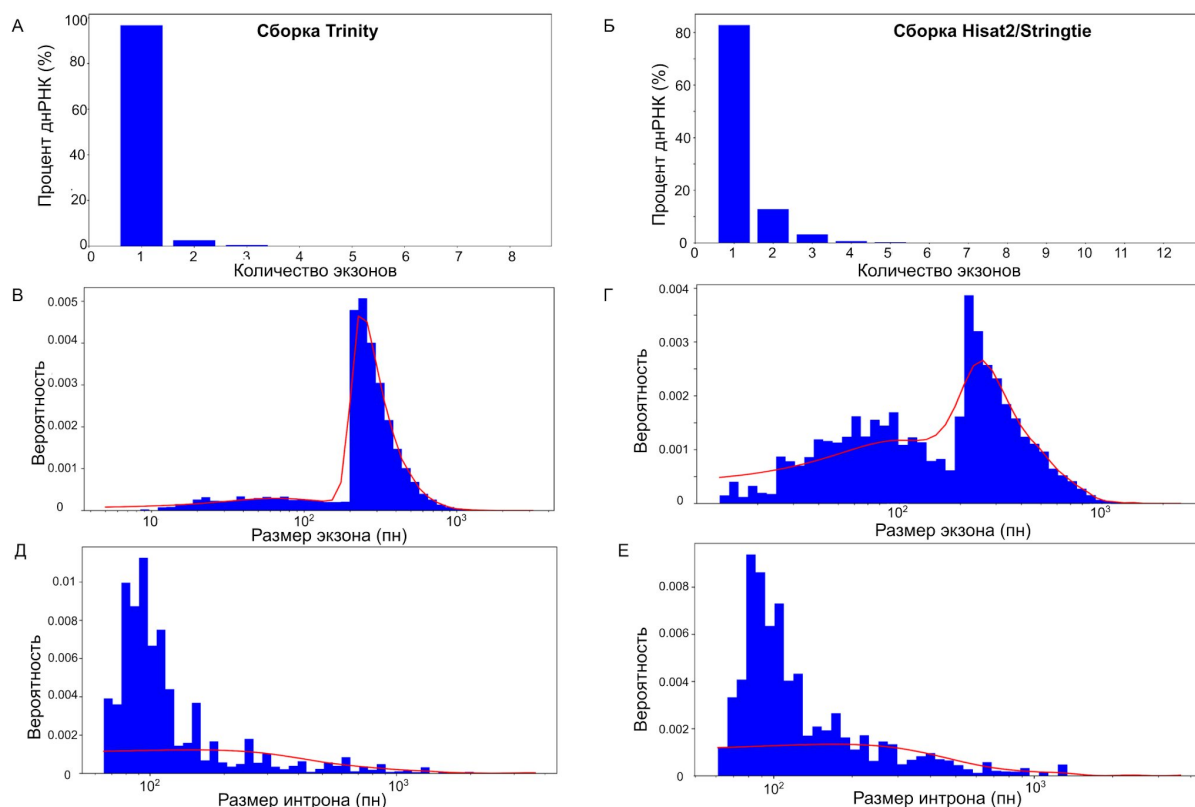


Рисунок 3.3. Анализ структурных характеристик днРНК. А,Б) Распределение количества экзонов приходящихся на одну днРНК В,Г) Распределение размера экзонов, Д,Е) Распределение размера интронов. Структурные характеристики днРНК полученных из сборки Trinity представлены на графиках А,В,Д. Структурные характеристики днРНК из сборки Hisat2/StringTie - Б,Г,Е.

Структурные характеристики известных днРНК из базы данных Ensembl Plants и выявленных днРНК представлены в таблице 3.4. Из таблицы видно, что пики распределения длин экзонов и интронов днРНК близки для сборок Trinity и Hisat2/StringTie. Эти характеристики представлены средним, минимальным и медианным значениями длин экзонов и интронов.

Однако наблюдаются отличия между структурными характеристиками двух сборок. Средняя длина транскрипта почти в полтора раза больше для сборки Hisat2/StringTie. Максимальная длина экзонов для сборки Trinity также в полтора раза больше максимальной длины экзонов для сборки Hisat2/StringTie.

Таблица 3.4 также позволяет сравнить структурные характеристики известных днРНК с характеристиками выявленных днРНК. Медианные и средние значения длин экзонов схожи между известными и новыми. Исключением

является максимальная длина экзона, которая в 2.5 и 3.6 раза больше, чем для сборок Trinity и Hisat2/StringTie, соответственно. Длины транскриптов также имеют отличия. Длина транскриптов известных днРНК превышает длину выявленных днРНК примерно в 6 раз. Однако наиболее существенное отличие наблюдается для длин интронов известных днРНК, которые на порядок превышают (максимальная длина интрона) длину интронов выявленных днРНК для обеих сборок, за исключением минимальной длины интронов, которая соизмерима с длинами интронов выявленных днРНК.

Таблица 3.4: Сравнение структурных особенностей известных днРНК кукурузы и предсказанных днРНК

	Известные днРНК	Новые днРНК	
		Trinity	Hisat2/StringTie
Средняя длина экзона	371	349	355
Максимальная длина экзона	8785	3545	2437
Минимальная длина экзона	2	9	13
Медиана длины экзона	262	298	303
Средняя длина транскрипта	2326	381	524
Средняя длина интрона	2956	470	406
Максимальная длина интрона	121413	3653	3652
Минимальная длина интрона	67	65	61
Медиана длины интрона	522	156	190

Для выявленных антисмысловых днРНК мы дополнительно рассмотрели такую характеристику, как порядковый номер экзона, на который происходит выравнивание в белок кодирующем гене (рисунок 3.4). Доля выровненных

днРНК резко уменьшается для каждого последующего экзона. Более 35% днРНК выравниваются на первый экзон белок кодирующего гена для обеих сборок. Более 10% днРНК выравниваются на второй экзон. Для последующих экзонов доля снижается не так резко.

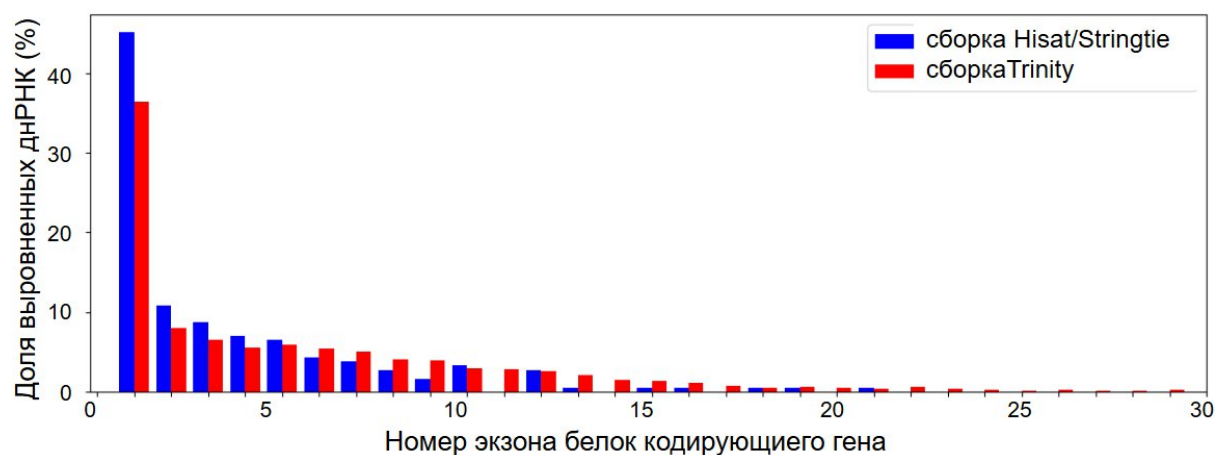


Рисунок 3.4. Распределение выравнивания антисмысловых днРНК в структуре гена-мишени по количеству покрываемых выравниванием экзонов.

3.1.5 Анализ видоспецифичности днРНК

Мы сравнили последовательности транскриптов выявленных днРНК с последовательностями известных днРНК из других организмов, представленных в базах данных днРНК, перечисленных в пункте 2.1.3. Высокая внутривидовая специфичность последовательностей днРНК подтверждается результатами кластеризации полного набора выявленных днРНК с последовательностями из внешних баз данных с помощью программы uCLUST (порог идентичности 50%). Это дает более 240 000 кластеров для сборок Trinity и Hisat2/StringTie. Большинство кластеров содержало последовательности только одного источника: либо известные днРНК из баз данных, либо днРНК, выявленные в настоящей работе. Медианный размер таких кластеров составлял 1-2 последовательности. Однако небольшая доля кластеров (1%) объединяла одновременно и выявленные, и известные днРНК. В этих кластерах среднее число последовательностей выше четырёх. Такая структура кластеризации подтверждает высокую внутривидовую специфичность днРНК: большинство из

них не находят гомологов среди аннотированных последовательностей других организмов.

Результаты, представленные в таблице 3.5, показывают высокую специфичность выявленных днРНК для кукурузы.

Таблица 3.5 Сходство выявленных днРНК кукурузы и последовательностей днРНК из баз данных

Характеристики сходства последовательностей	Сборка Trinity	Сборка Hisat2/StringTie
Общее количество выявленных днРНК, имеющих гомологию с днРНК из баз данных	8808 (14%)	1136 (31%)
Количество выявленных днРНК, имеющих гомологию только с днРНК кукурузы (внутривидовые)	8165 (13%)	1081 (30%)
Количество выявленных днРНК, имеющих гомологию только с днРНК других видов (поливидовые)	643 (1%)	55 (1%)

Большая часть (86%) выявленных днРНК не имеют гомологов с последовательностями днРНК из баз данных. Лишь 14% выявленных днРНК для Trinity и 30% выявленных днРНК для Hisat2/StringTie, имеет гомологию с последовательностями днРНК кукурузы из баз данных. Доля последовательностей, гомологичных днРНК из видов, отличных от кукурузы, составляет около 1% (таблица 3.5). Также стоит отметить, что доля выявленных днРНК, имеющих гомологию с другими видами, резко уменьшается с увеличением эволюционного расстояния от кукурузы: ~6% для *S. bicolor*, ~2% для *O. sativa*, ~0.1% для *B. distachyon*.

На рисунке 3.5 приведены распределения представленности классов днРНК кукурузы в зависимости от их близости к днРНК различных видов. Сборка

Trinity демонстрирует наибольшую долю антисмысловых днРНК для гомологичных последовательностей однодольных видов. Исключение составляет *A. thaliana* с более высокой долей гомологичных межгенных днРНК. Доля интронных днРНК, имеющих гомологию с другими видами, невелика и уменьшается с эволюционным расстоянием для других видов.

днРНК, полученные с помощью сборки Hisat2/StringTie, демонстрируют наибольшую долю межгенных днРНК, гомологичных последовательностям *A. thaliana*, *B. distachyon* и *Z. mays* из баз данных. Доля антисмысловых днРНК выше у гомологов *O. sativa* и *S. bicolor*. Доля интронных днРНК, гомологичных другим видам, невелика.

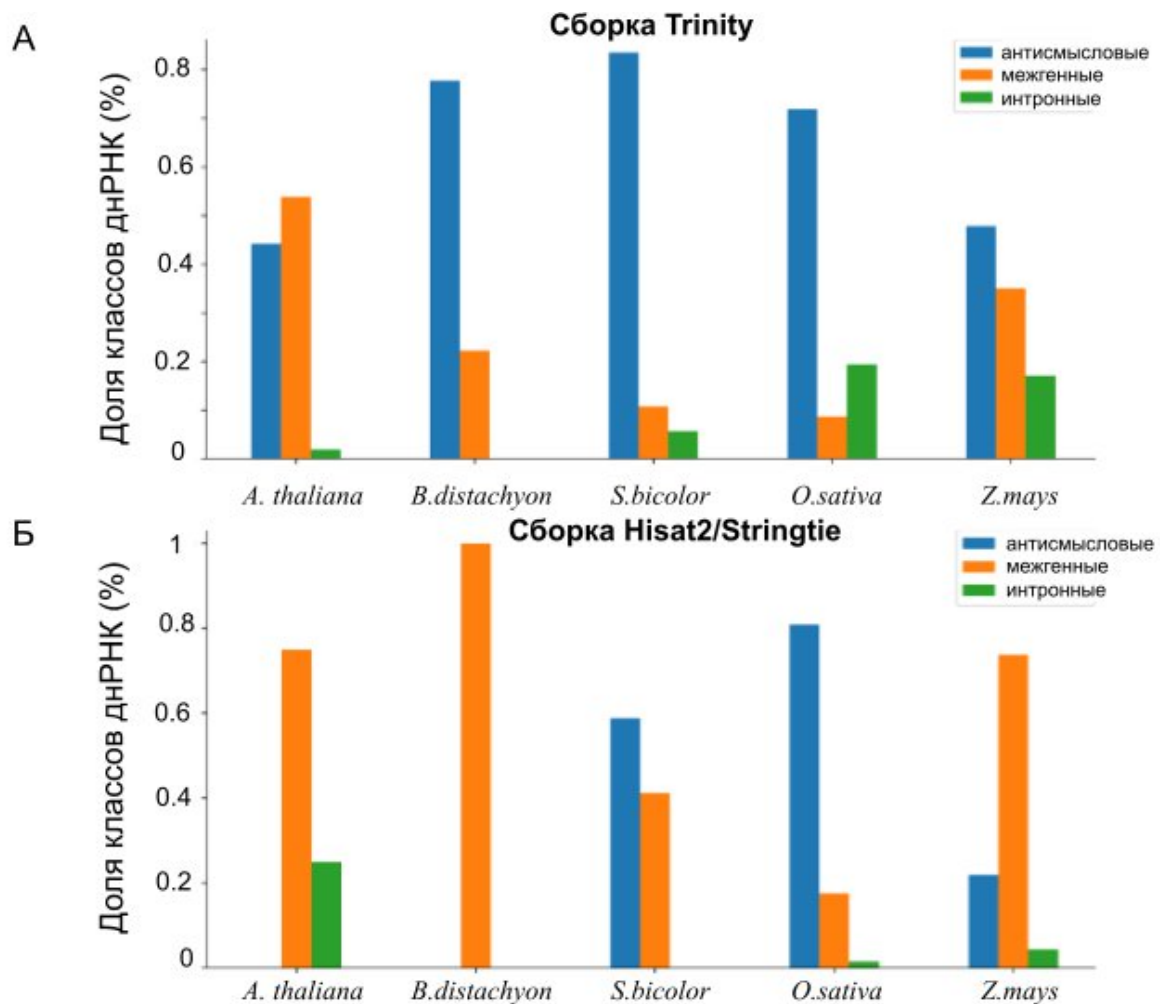


Рисунок 3.5. Распределение новых днРНК кукурузы по трем классам в зависимости от наличия гомологичных последовательностей у различных видов.

По оси X даны распределения для разных видов, по оси Y доля класса в %, цвет столбца зависит от класса днРНК и указан в верхнем левом углу диаграммы.

3.1.6 Анализ экспрессии днРНК в разных тканях кукурузы

Мы оценили распределение уровней экспрессии 2 выборок днРНК (поливидовые и внутривидовые) в различных тканях, рисунок 3.6.

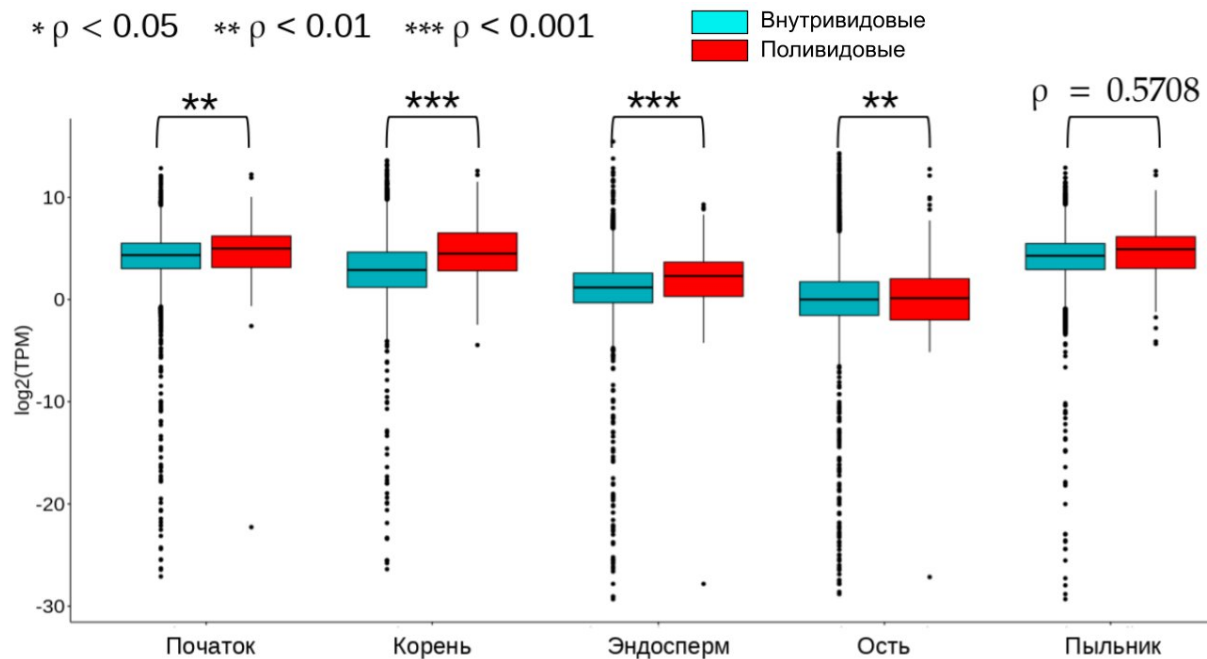


Рисунок 3.6. Анализ уровня экспрессии выявленных днРНК в тканях кукурузы. Диаграмма размаха для уровней экспрессии днРНК в различных тканях. Ось X - ткани, ось Y - нормированный уровень экспрессии днРНК в данных тканях.

Из рисунка 3.6 отмечается более высокий уровень экспрессии поливидовых днРНК в сравнении с внутривидовыми, в большинстве тканей. Кроме пыльника, где отличие уровней экспрессии днРНК незначительно. Большинство тканей на графике смещены в область больших значений TPM. Также стоит отметить, что у большинства тканей p -value меньше 0.05, что говорит о статистически значимых различиях между выборками. Однако для пыльника p -value > 0.05 , на основе чего можно сказать, что различия в экспрессии между выборками статистически не значимы.

Также мы оценили тканеспецифичность днРНК, рисунок 3.7. Мы провели этот анализ как для поливидовых, внутривидовых днРНК, а также для мРНК, идентифицированных в транскриптомах. В результате тканеспецифичность днРНК выше для ости (как для поливидовых, так и для внутривидовых днРНК). Эта тканеспецифичность как минимум в 10 раз выше, чем для других тканей. Интересно, что тот же паттерн тканеспецифичности между тканями мы наблюдали для мРНК.

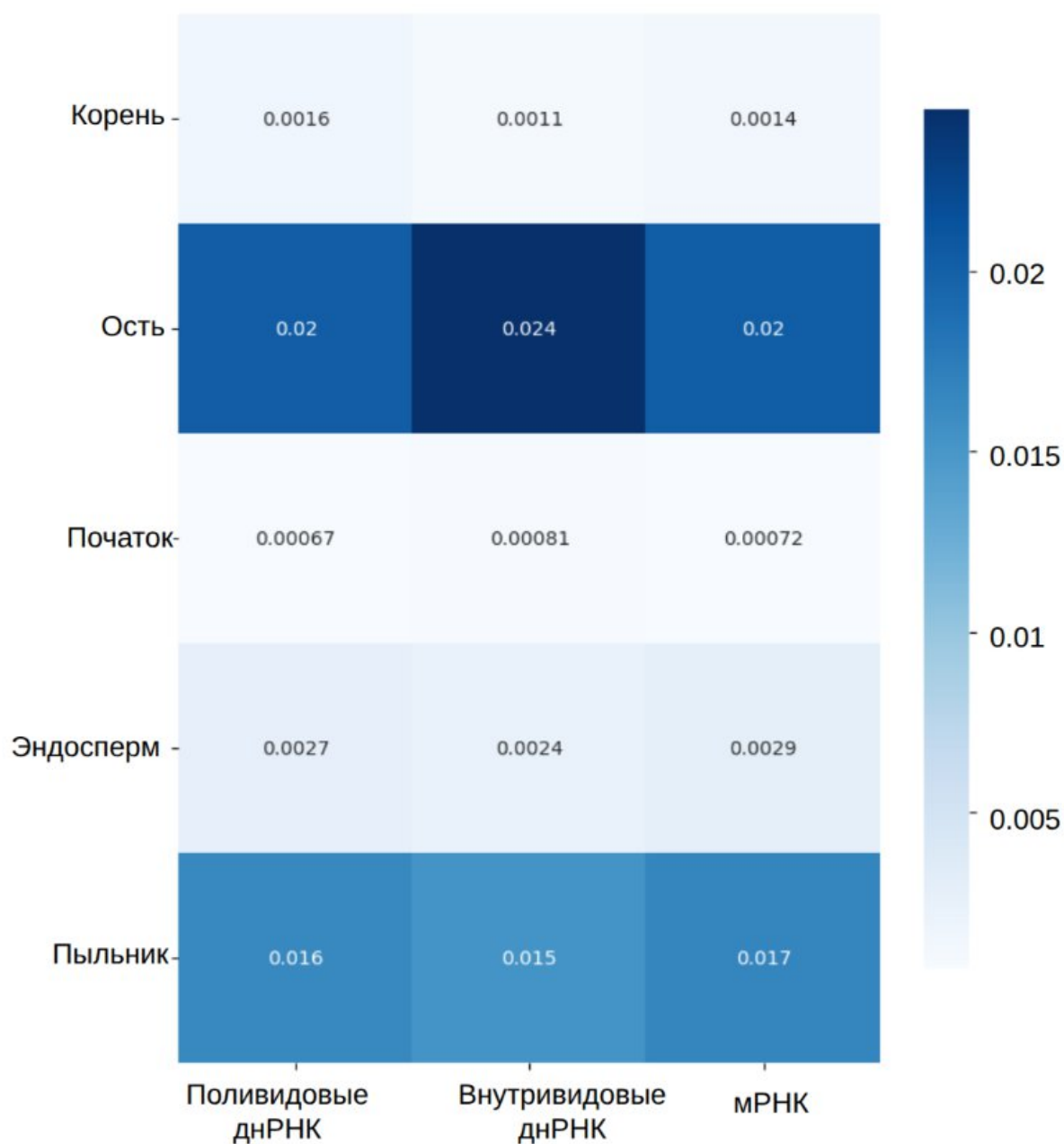


Рисунок 3.7. Тканеспецифичность днРНК кукурузы, представленная в виде тепловой диаграммы. По оси X представлены данные для двух групп днРНК

(поливидовые и внутривидовые) и мРНК. Соответствие цвета ячейки и величины тканеспецифичности показано шкалой справа от диаграммы.

3.2 Проверка работы разработанного конвейера ICApnoLncRNA на данных *A. thaliana*

Дополнительно мы использовали 840 библиотек *A. thaliana* в качестве модельного объекта для подтверждения корректности полученных результатов. Для ускорения расчетов было принято решение разделить наборы библиотек *A. thaliana* на 14 частей для последующей обработки. Мы получили 14 сборок *de novo* в программе Trinity, каждая из которых включает 60 библиотек. В общей сложности было получено 2 386 302 транскрипта. В результате работы биоинформатического конвейера ICApnoLncRNA было предсказано 70 263 длинных некодирующих РНК. Данные последовательности были классифицированы: 41 377 (58%) антисмысловых, 24 071 (34%) интронных, 4815 (6%) межгенных днРНК. Полученные доли классов днРНК для *A. thaliana* соответствуют результатам для сборки Trinity кукурузы (56%, 27%, 15% соответственно).

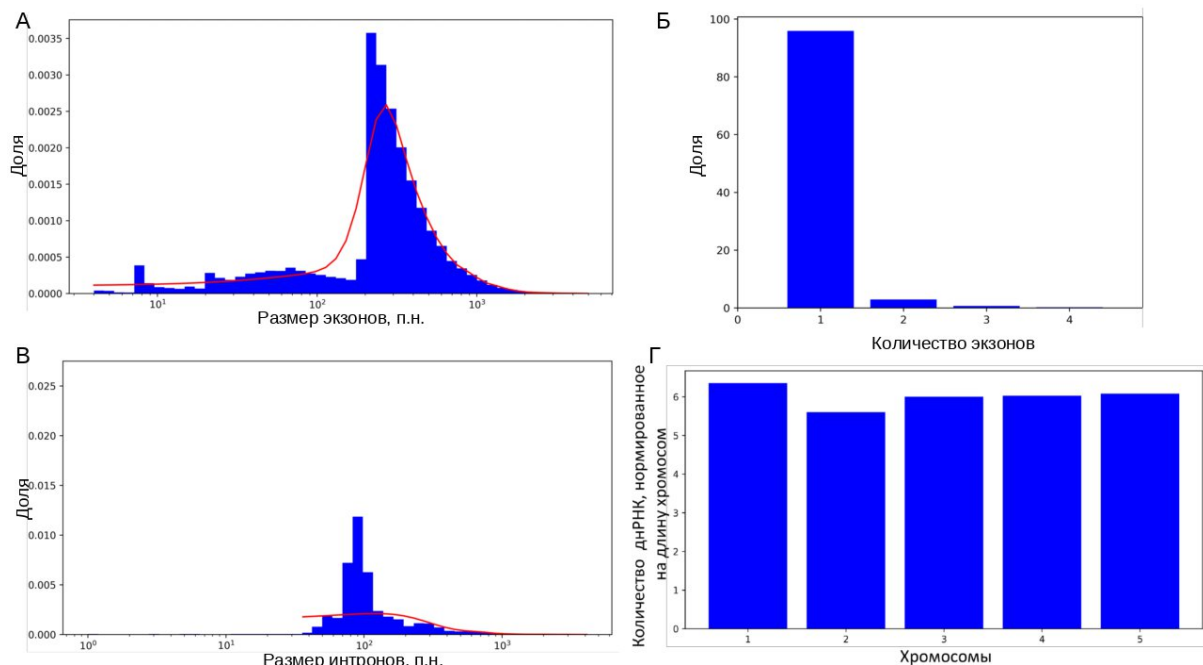


Рисунок 3.8. Структурные характеристики днРНК, полученных из сборки Trinity для *A. thaliana*. А) Распределение днРНК по размеру экзона. Б) Распределение

количества экзонов, приходящихся на одну днРНК. В) Распределение размера интронов. Г) Распределение днРНК по хромосомам.

Анализ структурных характеристик днРНК *A. thaliana* показал, что длина экзона днРНК находится в пределах 300-1000 п.н. (рисунок 3.8А). Также для днРНК было определено распределение по количеству экзонов (рисунок 3.8Б). На диаграмме видно, что большая часть (примерно 97%) днРНК имеет только один экзон. 3% днРНК имеют два экзона. Самая многочисленная группа интронов имеет размер около 80-200 п.н. (рисунок 3.8В). Стоит заметить, что существует совсем небольшая доля интронов, длина которых приближается к 1000 п.н. На последней диаграмме (рисунок 3.8Г) показано количество днРНК, нормированное на длину хромосом (1 днРНК на 10 000 п.н.). В целом можно сказать, что распределение днРНК по хромосомам арабидопсиса примерно равномерное. Больше всего днРНК располагается на первой хромосоме, а на второй - меньше по сравнению с остальными. Полученные структурные характеристики соответствуют как литературным данным для разных растений, так и полученным нами распределениям для кукурузы.

На последнем этапе работы конвейера мы проанализировали экспрессию днРНК в разных органах растения (рисунок 3.9). Исходя из рисунка 3.9, наибольшее значение экспрессии наблюдается в цветке как среди поливидовых, так и среди внутривидовых последовательностей днРНК. Менее всего днРНК экспрессируется в листе. Среди внутривидовых последовательностей корня, побега и пыльника нет существенных различий в экспрессии, в отличие от поливидовых днРНК в этих же частях растения. Так, в корне и побеге она выше, чем в пыльнике. В целом можно заметить, что экспрессия поливидовых днРНК выше, чем у внутривидовых. Полученные результаты экспрессии также находятся в согласии с результатами, полученными для кукурузы (см. раздел 3.1.6).

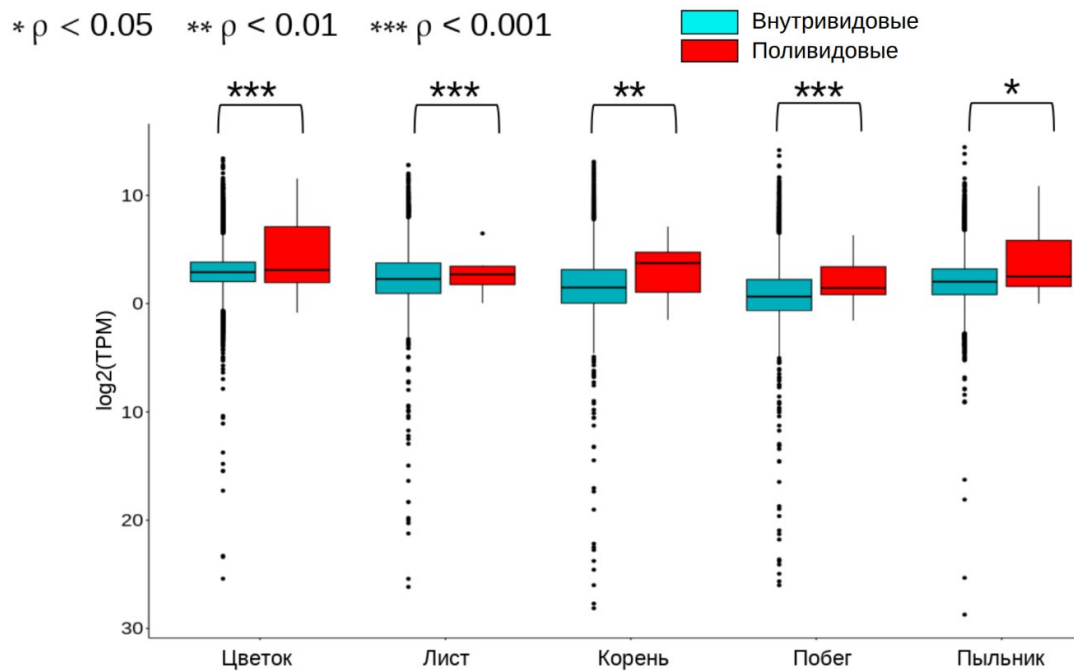


Рисунок 3.9. Анализ уровня экспрессии выявленных днРНК в тканях *A. thaliana*. Диаграмма размаха для уровней экспрессии днРНК в различных тканях. Ось X - ткани *A. thaliana*, ось Y - нормированный уровень экспрессии днРНК в данных тканях.

В результате мы показали сходство в распределении как структурных характеристик днРНК, так и экспрессии между *A. thaliana* и кукурузой. Это показывает, что, несмотря на высокую видоспецифичность днРНК, их структурные характеристики у разных организмов достаточно схожи.

3.3 Обсуждение

3.3.1 Сравнение характеристик разработанного конвейера ICApnoLncRNA с существующими

Существующие конвейеры идентификации и аннотации днРНК, такие как lncEvo (Bryzghalov et al., 2021), LncPipe (Zhao et al., 2018) или CALINCA (Talyan et al., 2021), разрабатывались для анализа транскриптома человека и животных. В основе работы всех конвейеров для некодирующих РНК лежит их идентификация и структурно-функциональный анализ. Для идентификации днРНК в большинстве перечисленных конвейеров используется программа CPC2 (Kang et al., 2017), которая разделяет выборку исследуемых транскриптов на кодирующие и некодирующие на основе четырёх признаков, связанных с

открытой рамкой считывания. Однако при распознавании некодирующих последовательностей эта программа демонстрирует достаточно высокий уровень ошибок (Yang, 2018; Li et al., 2014; Wang, 2013). В нашей работе мы использовали программу LncFinder, которая обладает более высокой точностью по сравнению с CPC2 Han et al., 2019. Кроме того, преимуществом LncFinder является то, что она основана на алгоритмах машинного обучения, параметры которых можно подобрать индивидуально для специфического референсного генома. Конвейер ICAAnnoLncRNA реализует такую возможность, используя данные по аннотированному днРНК референсного генома. Это позволяет достичь большей точности при идентификации днРНК у разных видов растений, структура геномов которых отличается большим разнообразием (Bolser et al., 2016). Отметим, что в отличие от методов глубокого обучения (Yang, 2018), алгоритм LncFinder не требует очень больших выборок (более 10 тыс. транскриптов каждой группы) для обучения. Это полезно в случае, когда построение метода распознавания основано на данных аннотации референсного генома, содержащего ограниченное количество известных днРНК (даже для хорошо изученного генома кукурузы это число составляет около 2 000). В процессе подбора параметров LncFinder оценивает точность распознавания днРНК; для исследованного нами генома кукурузы она оказалась высокой ($F1 = 0.951$ и $MCC = 0.86$).

Фильтрация потенциальных днРНК является необходимым этапом в любом конвейере, так как современные методы не идеальны и неизбежно содержат ошибки. LncPipe использует дополнительный фильтр по длине транскрипта (больше 200 п.н.) и по количеству экзонов (больше двух). CALINCA и lncEvo применяют схожие критерии фильтрации: длина последовательности (> 200 п.н.), уровень экспрессии ($FPKM > 1$, $TPM > 1$ соответственно), длина открытой рамки считывания не должна превышать 50 (CALINCA) и 100 (lncEvo) аминокислотных остатков. В нашем случае мы используем другую систему фильтров. Конвейер удаляет из исследования последовательности, экспрессирующиеся из ТЕ-областей, имеющие копии небольшого размера (возможный шум), а также не соответствующие классификации днРНК.

Картирование последовательностей транскриптов на референсный геном и сравнение результата с аннотацией известных генов (как кодирующих белки, так и некодирующих). Эту задачу мы решали с помощью программы *gffcompare*, которая активно используется в задачах идентификации днРНК (Halstead et al., 2021; Megquier et al., 2019; Singh et al., 2021; Muthaiyan et al., 2021; Azlan et al., 2019). Это позволяет отбросить транскрипты, которые полностью совпали с генами известных белков (результат неточного предсказания кандидатов днРНК). Следующая большая группа последовательностей, которая идентифицируется на основе этого метода: фрагменты, перекрывающиеся в положительном направлении с экзонами белок кодирующих генов. В некоторых работах такие последовательности рассматриваются как кандидаты в днРНК (Joshi et al., 2016; Bryzghalov et al., 2021). Однако ряд авторов полагает, что такие фрагменты могут соответствовать новым копиям известных генов (Halstead et al., 2021; Azlan et al., 2019). В нашем конвейере мы сделали выбор в пользу более строгого подхода к идентификации днРНК и исключили транскрипты этого типа из анализа. Таким образом, мы предлагаем пользователю наиболее надежный набор последовательностей транскриптов, предположительно кодирующих днРНК.

Как и в большинстве других работ, мы рассматриваем три класса выявленных днРНК: межгенные, интронные и антисмысловые. В рассмотренных выше биоинформатических конвейерах для классификации днРНК применяется программа *cuffcompare* (Trapnell et al., 2012), которая является аналогом *gffcompare*.

После выявления днРНК большинство конвейеров предлагают пользователю дополнительный анализ транскриптов. *lncEvo* предлагает анализ консервативности между двумя исследуемыми видами. *CALINCA* позволяет провести анализ дифференциальной экспрессии, экспрессии днРНК в разных тканях, консервативности с другими организмами. *LncPipe* - анализ дифференциальной экспрессии. Наш конвейер реализует большинство из перечисленных возможностей, а также позволяет анализировать структурные особенности днРНК (экзон-интронную структуру, особенности антисмыслового выравнивания на белок кодирующие последовательности). При этом

пользователю предоставляется возможность использовать для анализа собственные данные как по экспрессии, так и по последовательностям днРНК из различных баз данных.

Наш конвейер использует в качестве входных данных набор транскриптомных последовательностей, предоставленных пользователем. Таким образом, можно выбрать любой способ сборки, подходящий для задачи пользователя. Некоторых пользователей может интересовать сборка на основе референсного генома. Другие могут предпочесть сборку *de novo* даже в том случае, когда референсный геном известен и аннотирован. Например, создание сборок транскриптома *de novo* для модельных растений, таких как арабидопсис, рис и кукуруза, может быть полезно для обнаружения новых копий существующих аннотированных генов (Geniza et al., 2017).

3.3.2 Оценка точности идентификации днРНК и сравнение их структурных характеристик с существующими данными

Оценка качества идентификации днРНК

Оценка качества идентификации и анализа днРНК с помощью конвейера ICAnnoLncRNA производилась на основе следующих критериев: точность предсказания днРНК с помощью обученной модели, доля известных днРНК из баз данных, обнаруженных в наборе выявленных днРНК.

Оценка точности предсказания днРНК была произведена для модели, обученной на кукурузе. Обучение производилось на основе известных последовательностей мРНК и днРНК, с помощью двух метрик *F1* и *MCC*. В результате значение *MCC* = 0.86. Высокое значение *MCC* свидетельствует о сильной корреляции между предсказанными и фактическими классами транскриптов. Важно отметить, что *MCC* учитывает все четыре категории матрицы ошибок, что подтверждает высокую надежность модели и исключает вероятность того, что качество классификации является артефактом дисбаланса между количеством мРНК и днРНК в выборке. В свою очередь, значение метрики *F1* = 0.951 указывает на баланс между чувствительностью и точностью: модель

одинаково хорошо избегает как пропуска реальных днРНК (высокая полнота), так и ошибочного отнесения мРНК к некодирующим (высокая точность).

В результате проведенного анализа для кукурузы было выявлено 61 004 (52 526 локусов) и 3706 (3583 локусов) транскриптов днРНК для сборки Trinity и Hisat2/StringTie соответственно. Из них 306 (Trinity) и 297 (Hisat2/StringTie) транскриптов, полностью совпали с известными днРНК кукурузы (таблица 3.3). Референсный геном кукурузы B73 содержит 2535 аннотированных днРНК. Таким образом, нам удалось обнаружить 12% от этого набора. Данный результат связан с тем, что мы работали с 15 транскриптомными библиотеками, полученными из ограниченного набора тканей (см. таблица ПЗ), тогда как полный аннотированный набор днРНК был собран по результатам сотен экспериментов, охватывающих десятки тканей и стадий развития (Bolser et al., 2016). Помимо этого, экспрессия днРНК, как было показано в нашей работе и литературе, тканеспецифична (Derrien et al., 2012; Li et al., 2014), поэтому часть аннотированных днРНК не экспрессируется в исследованных тканях и вследствие этого не обнаруживается. Поэтому обнаружение 12% известных днРНК мы считаем удовлетворительным результатом выделения реальных некодирующих РНК, а не артефактов сборки или транскрипционного шума.

Сравнение структурных характеристик днРНК с существующими данными

Сравнение характеристик днРНК для рассматриваемых сборок продемонстрировало следующие сходства и различия. Обе сборки продемонстрировали общие черты в распределении локусов днРНК по хромосомам: частоты встречаемости выше для терминальных и ниже для центромерных участков хромосом. Это выглядит разумным из-за высокой доли идентифицированных нами антисмысловых и интронных днРНК (50 % и выше для обеих сборок). Эти днРНК пересекаются с генами, кодирующими белок, ближе к концам хромосом (Jiao, 2017). Наши результаты отличаются от ранее опубликованных работ (Li et al., 2014; Han et al., 2019), в которых распределение днРНК вдоль хромосом практически однородно. Данные различия могут возникать из-за того, что мы проводили дополнительную фильтрацию

транскриптов, связанных с мобильными элементами, которые имеют высокую частоту локализации в центромерных районах хромосом (Jiao, 2017).

Количество транскриптов, экспрессируемых на локус днРНК (~1), в обеих сборках также равно. Это значение близко к оценкам из других работ по поиску и анализу днРНК у кукурузы (Li et al., 2014), *A. thaliana* (Zhao, 2018) (~1.6 транскриптов на локус) и томата (Wang, 2015) (~1.6 транскриптов на локус). Следует также отметить, что днРНК, имеющие один экзон, доминируют в обоих типах сборок. Эта характерная черта была обнаружена и ранее в работе (Li et al., 2014).

Стоит также отметить, что средние/медианные значения распределений длин экзонов и интронов также близки как между обеими сборками кукурузы (~250 п.н. для экзонов и ~100 п.н. для интронов) так и со сборкой *A. thaliana*. Близки также оценки средней длины транскриптов днРНК (380-500 п.н.). Они близки к оценке, полученной (Li et al., 2014), - 463 п.н.

Оценка доли межгенных днРНК для исследованных нами транскриптомов как для кукурузы так и для *A. Thaliana* оказывается ниже, чем в других работах. Например, в работе (Li et al., 2014) 93% из 1704 соответствующих транскриптов классифицируются как межгенные днРНК и 7% - как антисмысловые. Эти авторы не исследовали интронные днРНК. В работе (Han et al., 2019) проанализировано 400 000 транскриптов и идентифицировано 18 000 днРНК. Эти авторы классифицировали 84% последовательностей как межгенные, 7% как антисмысловые и оставшиеся 3% как интронные. Исследование днРНК у *B. distachyon* и *B. napus* (De Quattro et al., 2017; Joshi et al., 2016) показывает преобладание межгенного класса днРНК - 81% и 88% соответственно. Подобно нашим результатам, анализ транскриптов днРНК у *Populus trichocarpa* и *A. thaliana* указывает на преобладание антисмыслового класса (Ye et al., 2022; Kindgren et al., 2020), интронный класс наименее распространен.

Поиск гомологичных днРНК среди известных последовательностей из баз данных показал, что лишь 1% выявленных в нашей работе днРНК имеет сходство с днРНК других видов растений. При этом наибольшее количество гомологов было обнаружено у однодольных растений *S. bicolor* и *O. sativa*. Данные

результаты схожи с результатами (Han et al., 2019), которые продемонстрировали, что только 2.5% и 0.1% последовательностей кукурузы гомологичны известным днРНК из сорго и риса соответственно. Наши результаты согласуются с выводами из предыдущих работ: последовательности днРНК в высокой степени внутривидовые, и их сходство с днРНК других видов резко падает с увеличением эволюционного расстояния (Deng et al., 2018; Sang et al., 2021). Гомологичные последовательности у других видов чаще встречаются для антисмысловых днРНК в сборке Trinity и для интронных днРНК в сборке Hisat2/StringTie. Это может объясняться более высокой встречаемостью этих классов днРНК в соответствующих сборках. Эти результаты согласуются с работой (Deng et al., 2018). Авторы сравнили днРНК, идентифицированные биоинформатическими методами у *A. thaliana* и *O. sativa*, с днРНК 10 видов растений и обнаружили, что чем ближе два вида находятся на филогенетическом дереве, тем больше у них сходных последовательностей днРНК. В целом, наш анализ показывает, что последовательности днРНК в основном внутривидовые. По сравнению с различными видами растений наибольшая доля наших транскриптов имеет сходство с уже известными днРНК из кукурузы.

Исследование экспрессии выявленных днРНК указывает на более высокую экспрессию поливидовых днРНК по сравнению с внутривидовыми днРНК. Кроме того, анализ тканеспецифичности указывает на высокую экспрессию поливидовых и внутривидовых днРНК в ости кукурузы. Однако нельзя напрямую сравнивать наши результаты с результатами (Han et al., 2019), которые оценили долю транскриптов разных типов в одних и тех же тканях. В нашей работе мы использовали дополнительную нормировку по количеству экспрессированных транскриптов во всех тканях, что, по нашему мнению, более адекватно для определения специфичности экспрессии.

ГЛАВА 4 Сборка и анализ пан-транскриптома инбредных линий кукурузы

4.1 Результаты

4.1.1 Сборка пан-транскриптома линий кукурузы

Дальнейшие разделы направлены на сборку пан-транскриптома кукурузы и анализ днРНК в данных масштабах. Различные оценки качества сборки *de novo* транскриптомов тремя методами представлены в таблице 4.1.

Таблица 4.1: Результаты *de novo* сборки 503 транскриптомов кукурузы 3 методами.

Показатели качества сборки транскриптома	TrinityGG	Trinity	SPAdes-rna
Средняя длина контига	942	864	592
Общее количество собранных транскриптов	882 370	21 646 755	44 149 357
GC%	47	51	51
N50	1834	1514	1475

Наибольшее количество транскриптов было получено для сборки SPAdes-rna - 44 149 357. Наименьшее - для сборки TrinityGG на основе референсного генома. Однако все три сборщика демонстрируют высокие показатели качества сборки, которые хорошо согласуются.

Результаты фильтрации транскриптов по длине (>200 п.н.) и классификации на кодирующие и не кодирующие для различных сборщиков представлены в таблице 4.2.

Таблица 4.2: Результаты разделения транскриптомов кукурузы на белок кодирующие гены и не кодирующие гены.

Этапы разделения	TrinityGG	Trinity	SPAdes-rna
Входные данные	896 138	21 646 755	44 149 357
Фильтр по длине последовательности	895 964 (99%)	21 646 755 (100%)	31 740 702 (71%)
Не кодирующие	647 415(73%)	12 951 969 (59%)	19 194 063 (60%)
Кодирующие	248 549 (27%)	8 694 786 (40%)	12 546 639 (40%)

Из таблицы 4.2 видно, что все три сборщика дают на выходе достаточно длинные последовательности: не менее 70% проходят фильтр по длине. При этом сборка SPAdes-rna дает больше коротких транскриптов, по сравнению с Trinity. Таблица также демонстрирует, что в транскриптомах преобладают не кодирующие РНК. Эти пропорции незначительно отличаются для разных сборщиков: для алгоритмов *de novo* доля не кодирующих последовательностей меньше, чем для алгоритма, основанного на референсном геноме.

Результаты применения последующих этапов обработки для транскриптов мРНК и днРНК приведены в таблицах П4.1 и П4.2 соответственно; для не кодирующих последовательностей детали приведены в таблицах П4.3 и П4.4.

Таблица 4.3: Результаты сборки пан-транскриптома кодирующей и не кодирующей частей

Этапы обработки	мРНК	днРНК
Мета-сборка	17 795 733	2 007 596
Удаление дубликаций	14 392 809	2 007 596
Кластеризация	245 436	595 198

Из таблицы П4.1 видно, что количество последовательностей белок кодирующих генов увеличилось после применения Evidential Gene, за счёт восстановления полноразмерных открытых рамок считывания, разделения химерных контигов на независимые локусы и сохранения альтернативных изоформ в рамках каждого генного кластера. Данная программа позволяет предсказывать возможные альтернативные (неканонические) стартовые кодоны и проверять их наличие в последовательности. Авторы рекомендуют оставлять оба варианта (Gilbert, 2019). Таким образом, мы оставили оба варианта последовательностей, полученных используемой программой.

Полученные в результате кластеризации центроиды мы считаем пан-транскриптомами для кодирующей и не кодирующей частей.

4.1.2 Проверка качества сборки транскриптомов

Чтобы убедиться в надёжности исходных данных для последующего анализа, мы оценили качество сборки транскриптома, полученного с помощью конвейера PTA для референсной линии кукурузы. Данный подход был выбран потому, что в исходной работе (Hirsch et al., 2014) индивидуальные геномы линий не собирали, и единственным эталоном остаётся геном В73. Результаты проверки качества сборки представлены в таблице 4.4 и на рисунке 4.1.

Таблица 4.4: Параметры оценки качества мета-сборки для референсной линии кукурузы В73.

Параметры	днРНК	мРНК
Общее количество транскриптов	39 315	25 851
Картированные	26 342 (67%)	25 829 (98%)
Некартированные	12 973 (33%)	22 (<1%)
Средняя доля картированных фрагментов	0.99	0.981
Ошибки сборки	365	1029
Покрытие аннотированных генов	0.011	0.078

Исходя из полученных результатов, более 98% белок кодирующих генов картируются на референсный геном. При этом средняя доля картированных фрагментов составила 0.981. Это означает, что практически все транскрипты мРНК, собранные конвейером, являются результатом экспрессии генов В73, а не являются случайными артефактами. Доля картированных днРНК ниже и составляет 67%. Существенную часть невыравненных днРНК можно объяснить тем, что это результат экспрессии генов, которые не представлены в финальной сборке хромосом генома В73. Однако именно они могут составлять наиболее интересную, вариабельную часть не кодирующего пан-транскриптома. При этом доля картированных фрагментов для днРНК составила 0.99. Стоит отметить, что доля ошибок сборки у днРНК ниже, чем у белок кодирующих генов. Это может быть связано с тем, что для фильтрации днРНК мы применяли больше различных параметров, благодаря которым повысилось качество сборки.

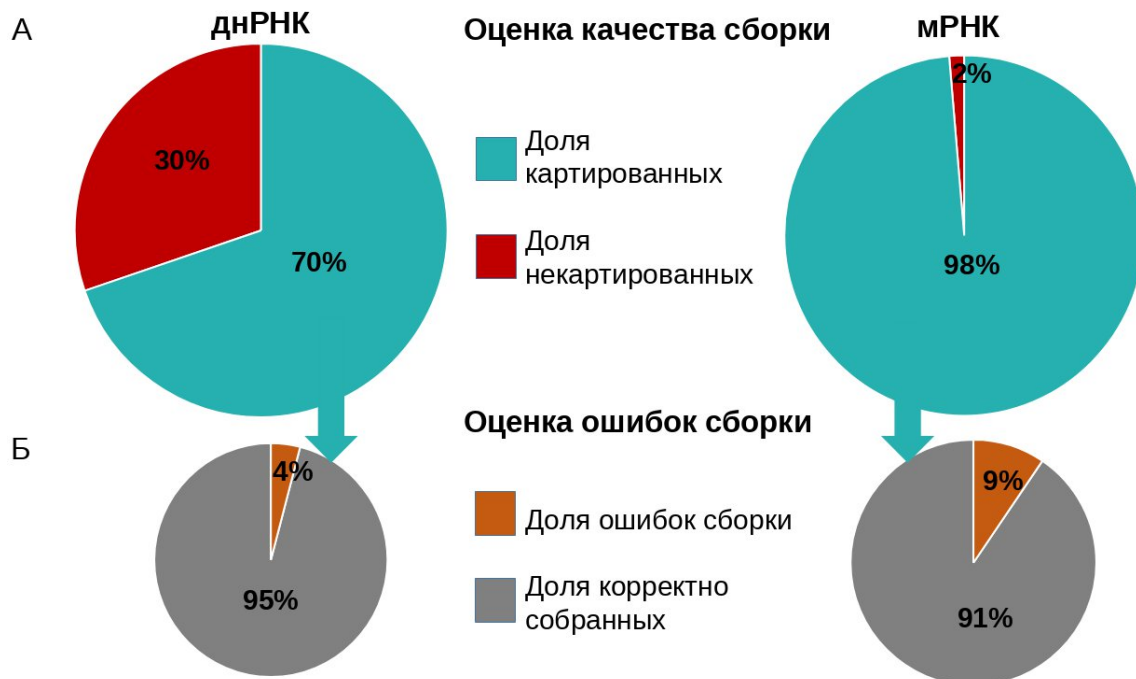


Рисунок 4.1. Оценка качества сборки для мета-сборки белок кодирующих генов и днРНК. А) Оценка качества выравнивания. Б) Оценка ошибок сборки для картированных транскриптов; показана доля транскриптов, собранных некорректно.

В отличие от остальных параметров оценки качества, покрытие аннотированных генов имеет невысокие значения. Однако нужно учитывать, что это оценка качества для одной библиотеки из 503. То есть в единственном образце принадлежащем линии В73, экспрессируется лишь ограниченный набор генов (7.8% для белок кодирующих генов и 1.1% для днРНК), тогда как полная аннотация референсного генома описывает гены, работающие в самых разных органах, на разных стадиях развития и в ответ на различные стимулы.

Описанные выше результаты демонстрируют высокое качество сборки транскриптомов с использованием нашего конвейера. Это позволяет с высокой долей уверенности говорить, что полученные нами транскрипты белок кодирующих генов и днРНК не являются ошибками сборки и на их основе можно проводить дальнейший анализ пан-транскриптомов.

4.1.3 Анализ структуры пан-транскриптомов

Результаты анализа структуры пан-транскриптомов для кодирующих и некодирующих последовательностей представлены на рисунке 4.2. Из рисунка видно, что значения числа транскриптов в обоих пан-транскриптомах выходят на плато по мере роста числа анализируемых инбредных линий. Таким образом, они являются закрытыми, о чем свидетельствуют и значения параметра $\alpha_{pan} > 1$ для степенного закона распределения. Тем не менее между двумя пан-транскриптомами обнаруживаются и заметные различия. Анализ зависимости числа уникальных транскриптов от количества добавляемых транскриптомов показал, что для охвата 99% общего набора белок кодирующих генов пан-транскриптома кукурузы достаточно выборки из 93 инбредных линий. Для днРНК размер пан-транскриптома в зависимости от числа линий растет медленнее и охват полного набора днРНК происходит при числе линий, равном 395. Эти различия отражены и в значениях параметра α_{pan} : для мРНК его значения выше (1.70), чем для днРНК (1.33).

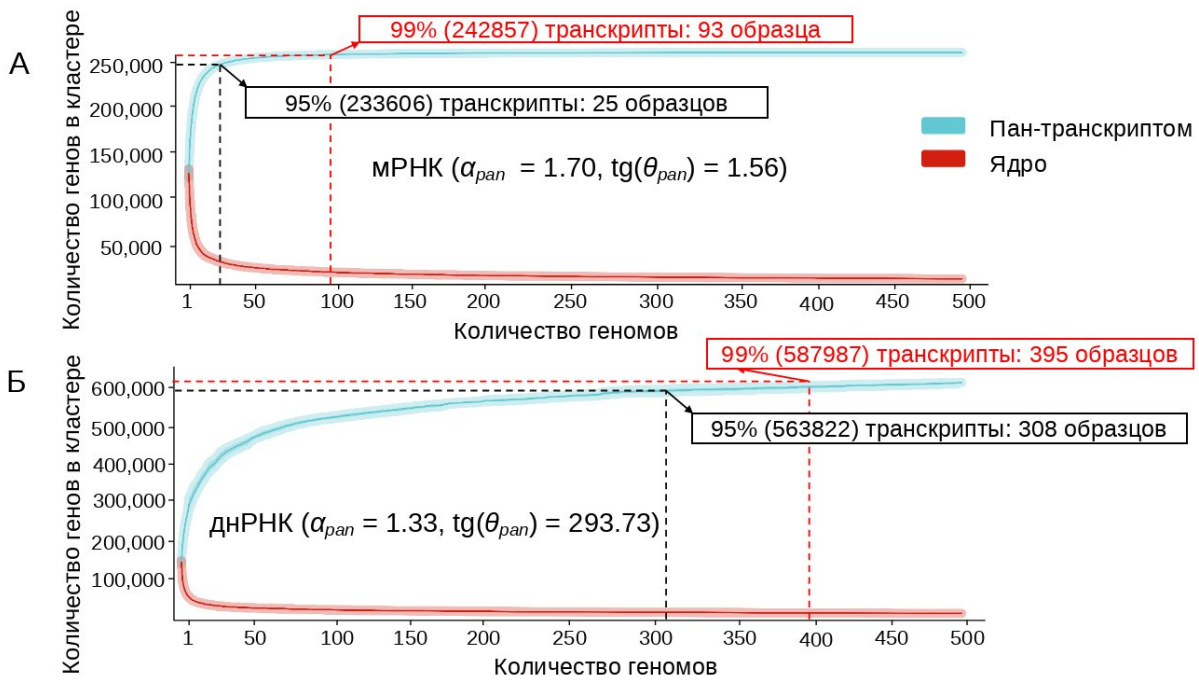


Рисунок 4.2. Структуры пан-транскриптомов инбредных линий кукурузы для кодирующих А) и не кодирующих Б) последовательностей. По оси X отложено число инбредных линий кукурузы. По оси Y отложено количество кластеров последовательностей, формирующих пан-транскриптом (голубая линия) и его ядро (красная линия). На графиках в скобках приведены параметры степенного распределения для соответствующих пан-геномов.

Отметим также, что при более медленном росте пан-транскриптома днРНК размер его ядерной части уменьшается быстрее, чем для мРНК: уже после добавления 30-й линии он практически не меняется, стремясь к значению 27 582 кластеров (рисунок 4.2). Для мРНК изменение размера ядра прекращается при добавлении ~60 линий кукурузы, приближаясь к значению 19 798 кластеров. Эти результаты свидетельствуют о более высоком разнообразии пан-транскриптома днРНК по сравнению с мРНК.

Результаты распределения транскриптов по доле линий, в транскриптомах которых они обнаруживаются, представлены на рисунке 4.3 и позволяют выделить для пан-транскриптомов не только ядерную часть, но также оболочку и облако. Для пан-транскриптома мРНК ядерная часть включает 8% генов, а большая часть генов относится к оболочке. При этом по мере уменьшения числа

линий, для которых они являются общими, доля мРНК растет (рисунок 4.3А). Максимальное значение (~25%) наблюдается для 10% линий. Однако при еще большем уменьшении доли общих линий (<10%, облако) доля мРНК уменьшается почти в 2 раза (до ~12%).

Диаграмма распределения транскриптов по линиям для днРНК существенно отличается. Она демонстрирует, что по мере уменьшения числа линий кукурузы доля общих днРНК монотонно растет, достигая максимума при 10% линий (52% генов). При этом доля генов днРНК, представленных в оболочке и ядре оказываются почти в 2 раза меньше, чем у мРНК (42% против 81% и 5% против 8%; рисунок 4.3В,Г). Таким образом, эти данные свидетельствуют о более высоком разнообразии последовательностей из пан-транскриптома днРНК по сравнению с мРНК.

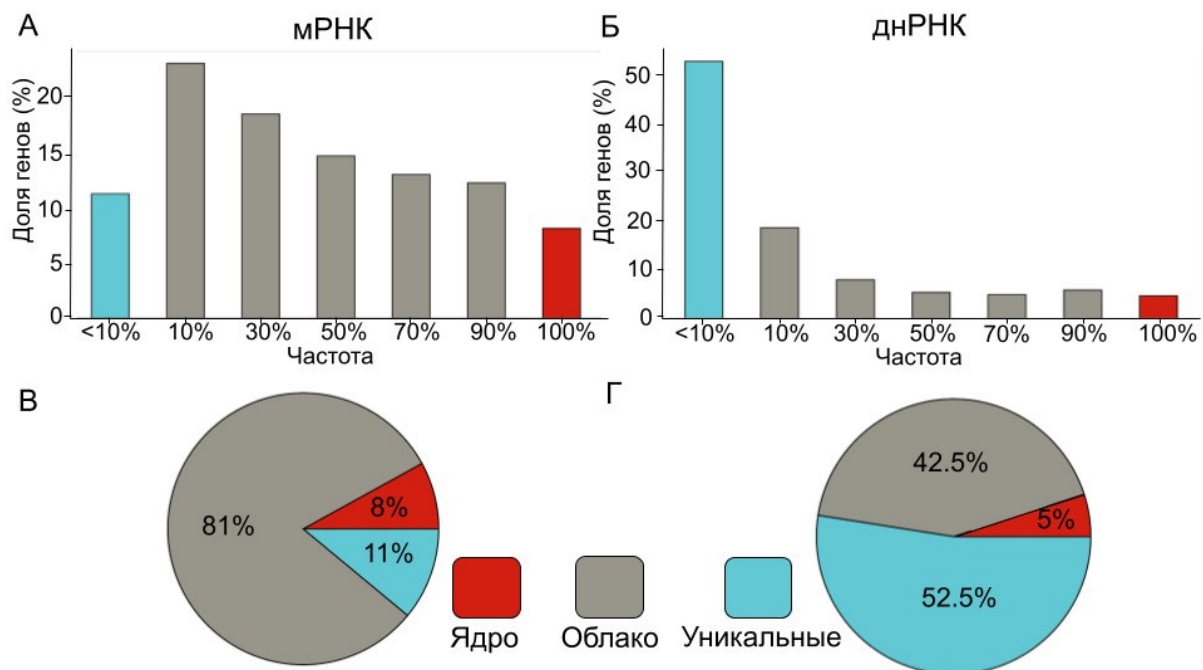


Рисунок 4.3. А,Б) Распределение генов в транскриптомах кодирующих А) и не кодирующих Б) РНК по доле инбредных линий кукурузы, для которых они являются общими. По оси Х отложены доли линий кукурузы от общего количества, по оси Y - доля РНК, представленных в этих линиях. В,Г) Распределение генов, представленных в ядре, оболочке и облаке пан-транскриптома для кодирующих В) и не кодирующих Г) РНК.

Для индивидуальных транскриптомов инбредных линий доля генов, входящих в ядро, составляет приблизительно одинаковую долю (~18%) как для мРНК, так и для днРНК (рисунок 4.4).

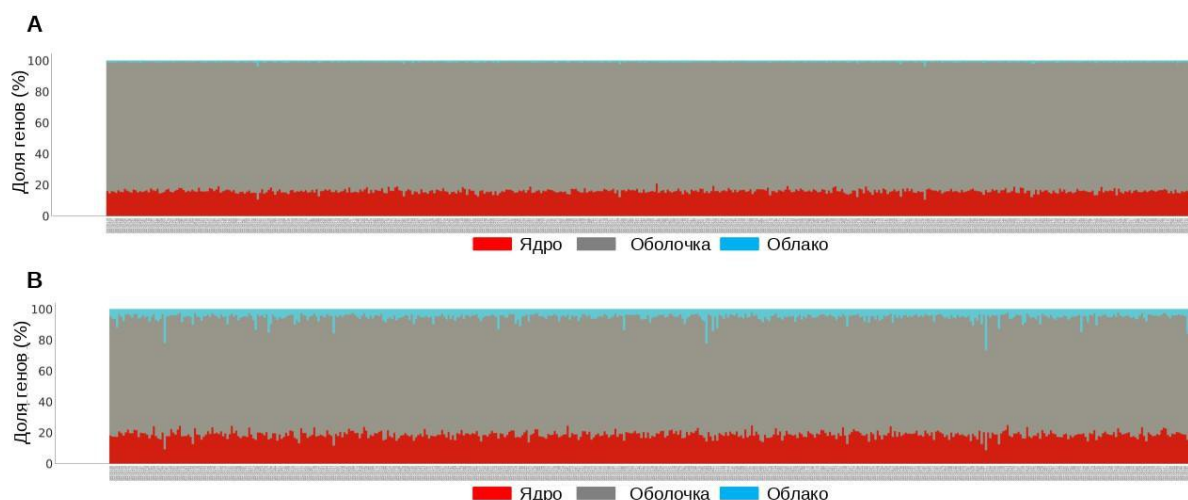


Рисунок 4.4. Гены, принадлежащие ядру, оболочке и облаку, в каждом транскриптоме для мРНК А) и днРНК Б) По оси X показаны 503 библиотеки кукурузы, по оси Y — доля генов, принадлежащих определённому типу в пан-транскриптоме.

Однако относительные доли генов, входящих в облако, различаются в несколько раз: в то время как у мРНК она составляет в среднем ~1-2%, для днРНК эта доля возрастает до 5-8%. Среднее значение по транскриптомам днРНК для каждой фракции составляет: ядро - 18%, оболочка - 75%, облако - 5%. Для белок кодирующих генов: ядро - 16%, оболочка - 83%, облако - 1%. У днРНК по сравнению с мРНК снижена доля генов, представленных в оболочке. Отметим, что на графиках (рисунок 4.4) наблюдаются выбросы - линии, которые заметно отличаются по долям ядро/оболочка/облако, при этом число таких линий больше для пан-транскриптома днРНК, что еще раз свидетельствует в пользу его большего разнообразия.

4.1.4 Анализ нуклеотидного разнообразия пан-транскриптомов

Поиск и анализ полиморфизмов в пан-транскриптоме кукурузы, состоящем из некодирующей части (днРНК) и кодирующей части (мРНК), первоначально

выявил 7 996 789 и 3 299 149 ОНП соответственно. После фильтрации по качеству (см. Материалы и методы, п. 2.2.5) для днРНК осталось 3116 наиболее значимых ОНП, а для мРНК - 2743 ОНП.

Результаты оценки нуклеотидного разнообразия π , плотности ОНП и распределения показателя Таджимы D приведены на рисунке 4.5. Средние значения нуклеотидного разнообразия составили для днРНК 3.677×10^{-5} , для мРНК 2.899×10^{-5} (рисунок 4.5А), что позволяет отвергнуть гипотезу об их равенстве ($p\text{-value} < 0.0001$, согласно критерию Уилкоксона). Дополнительная статистика распределений π для мРНК и днРНК представлена в таблице 4.5. Таблица показывает, что медиана π выше для днРНК (1.94×10^{-5}), чем для мРНК (1.58×10^{-5}). Распределение π для днРНК имеет более широкий диапазон по сравнению с мРНК (межквартильный размах 3.77×10^{-5} против 2.42×10^{-5}). Таким образом, последовательности днРНК демонстрируют большие вариации нуклеотидов по сравнению с мРНК.

Таблица 4.5: Дополнительные статистические данные для анализа значений π нуклеотидного разнообразия кодирующих и некодирующих РНК в пан-транскриптоме кукурузы.

Параметры	мРНК, π	днРНК, π
Медиана	1.58×10^{-5}	1.94×10^{-5}
Нижний квартиль Q1	8.31×10^{-6}	9.33×10^{-6}
Верхний квартиль Q3	3.26×10^{-5}	4.70×10^{-5}
Минимум	6.67×10^{-7}	9.00×10^{-7}
Максимум	0.09×10^{-3}	0.02×10^{-2}
Межквартильный размах	2.42×10^{-5}	3.77×10^{-5}
Стандартное отклонение	4.69×10^{-5}	4.46×10^{-5}

Сравнение гистограмм распределений плотности ОНП для кодирующих и некодирующих последовательностей показывает (рисунок 4.5Б), что для мРНК

это распределение смещено в сторону меньших значений (1 ОНП на 1 кб приходится в 60% случаев для мРНК и в 40% для днРНК).

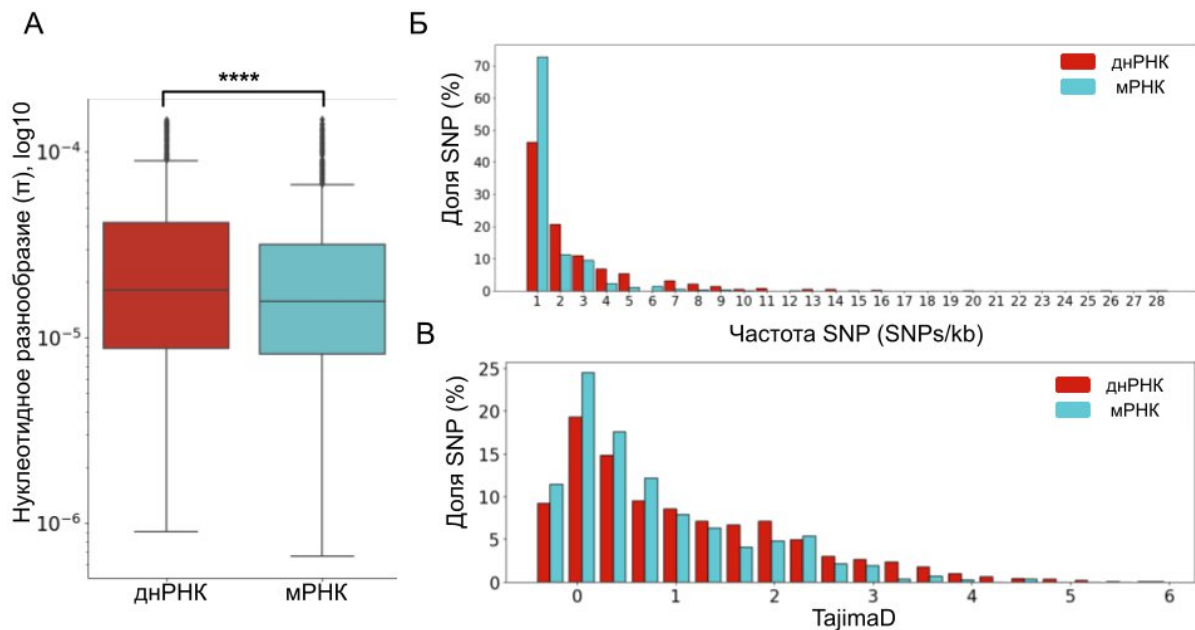


Рисунок 4.5. Анализ нуклеотидного разнообразия пан-транскриптомов кодирующих и некодирующих РНК кукурузы. А) Диаграмма размаха распределений нуклеотидного разнообразия (π) на 10 кб. Ось Y соответствует десятичному логарифму π . Б) Гистограммы распределения плотности ОНП. Ось X соответствует числу ОНП на 1 кб, Y - доле фрагментов. В) Гистограмма распределения показателя Таджимы D на 10 кб. Ось X соответствует значению D , ось Y - доле фрагментов. Диаграммы для днРНК показаны красным, для мРНК - голубым цветом.

Расчет показателя Таджимы D для обеих выборок показал, что его распределение смещено в область положительных значений (рисунок 4.5В). Доля отрицательных значений составляет 19% для мРНК и 17% для днРНК, при этом значения менее -0.44 не наблюдаются. Доля последовательностей, для которых $D > 2$ составила для мРНК 13%, для днРНК 20%. Эти данные свидетельствуют в пользу того, что для существенной доли генов, рассмотренных в данном анализе (как кодирующих, так и некодирующих), режим эволюции отличен от нейтрального, они содержат меньше редких аллелей.

4.1.5 Анализ синонимических и несинонимических замен

Для расчета скоростей нуклеотидных замен K_a и K_s для мРНК мы сформировали выборку из 72 539 референсных последовательностей CDS кукурузы. На CDS было выравнено 8 163 024 последовательностей мРНК пан-транскриптома, 45% от общего количества выявленных нами белок кодирующих генов (17 795 733) (таблица 4.3). Из них 4222 относились к ядру, 16 664 к оболочке, 1174 к облаку. В результате для белок кодирующих генов наблюдается 6 201 766 нуклеотидных замен.

Для оценки скорости K_n (нуклеотидных замен для днРНК) выборка включала 26 465 днРНК из БД RLncDB v.2.0. На референсные последовательности днРНК выравнено 150 592 днРНК пан-транскриптома, выборка включала 7% от общего количества выявленных днРНК, таблица 4.3. Из них 583 относились к ядру, 5436 к оболочке, 1492 к облаку. В результате для днРНК получено 144 580 замен.

Гистограммы распределений K_s , K_a и K_n для трех частей пан-транскриптома, а также оценки параметров α , β для аппроксимирующих их гамма-распределений представлены на рисунке 4.6А,Б,В. Из этих гистограмм видно, что меньшим значениям K_a , K_s соответствует большая высота столбцов, по сравнению с K_n . В то же время, для больших значений этих параметров столбцы распределения K_n оказываются выше. Это свидетельствует о более высоких скоростях замен нуклеотидов в днРНК по сравнению как с несинонимическими, так и синонимическими заменами мРНК. Эти закономерности хорошо прослеживаются при сравнении параметра α : он увеличивается в ряду $K_a < K_s < K_n$ систематически для всех трех частей пан-транскриптома. Это говорит о том, что темпы накопления нуклеотидных замен в днРНК выше, чем синонимических у мРНК; при этом для мРНК всегда выполняется неравенство $\alpha(K_a) < \alpha(K_s)$, что в целом характерно для замен мРНК в режиме стабилизирующего отбора.

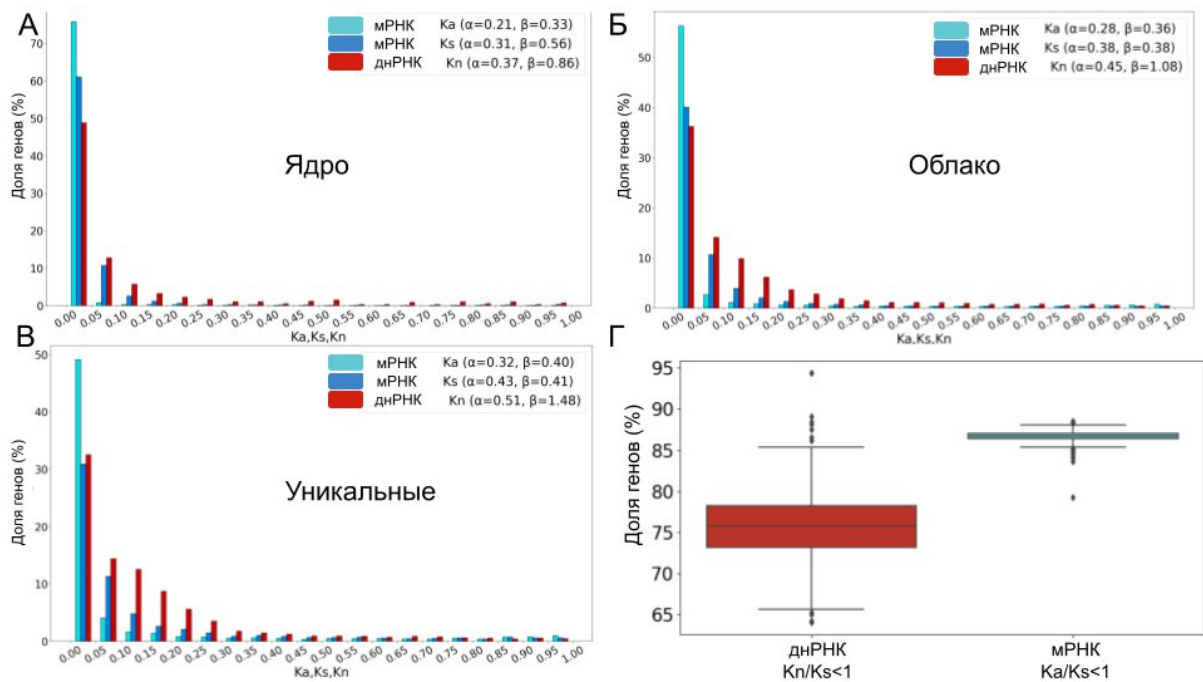


Рисунок 4.6. Анализ синонимических и несинонимических замен А,Б,В) Распределение типов генов ядра, оболочки, облака, относительно их значений K_n , K_s , K_a . По оси X направлены промежутки значений K_n , K_s , K_a , по оси Y доля генов, принадлежащих данному промежутку. Г) Распределение долей генов из фракции $K_a/K_s < 1$ и $K_n/K_s < 1$, для каждой библиотеки. По оси Y — доля генов, имеющих значение $K_a/K_s < 1$ и $K_n/K_s < 1$, составляют внутри каждой библиотеки.

Еще одной интересной закономерностью, которую можно наблюдать на рисунке 4.6А,Б,В, являются различия параметра α для различных частей пан-транскриптома. Для всех трех параметров K_s , K_a , K_n наблюдается его систематическое увеличение в ряду ядро<оболочка<облако. Это свидетельствует об ослаблении отбора для последовательностей соответствующих частей пан-генома как для мРНК, так и для днРНК.

Для генов, кодирующих мРНК в каждой линии кукурузы, мы оценили отношение параметров K_a/K_s , и долю консервативных генов, для которых $K_a/K_s < 1$. Аналогично мы рассчитали отношение K_n/K_s для каждого гена днРНК и долю консервативных генов с $K_n/K_s < 1$. В данном случае за величину K_s брали одинаковое для всех днРНК среднее для линии значение, как это рекомендовано разработчиками программы KaKs_calculator v3 (Zhang, 2022). Распределение

доли консервативных для пан-транскриптома днРНК и для пан-транскриптома мРНК представлено в виде диаграммы размаха на рисунке 4.6Г. Из рисунка видно, что среднее значение доли консервативных генов для днРНК (76%) существенно ниже, чем для кодирующих последовательностей (87%). Кроме того, для распределения доли консервативных генов днРНК характерен существенно больший размах, чем для распределения доли консервативных генов мРНК. Все это подчеркивает более высокую вариабельность днРНК кукурузы по сравнению с кодирующими белки и, соответственно, меньшее давление отбора на последовательности днРНК по сравнению с мРНК.

4.1.6 Разнообразие днРНК, различающихся по расположению в геноме

С помощью программы ICAAnnoLncRNA (Pronozin, 2023) для последовательностей днРНК, выравненных на референсный геном, мы провели классификацию по расположению: антисмысловые, межгенные и интронные. Для каждого типа днРНК был проведен анализ представленности в составе разных частей пан-транскриптома и скоростей замен K_n . Результаты представлены на рисунке 4.7.

Из рисунка видно, что антисмысловые днРНК наиболее часто представлены в ядре пан-транскриптома (~80%). В то же время для оболочки их доля снижается до 50%, а в облаке они представлены очень редко (~25%). Для межгенных и интронных днРНК наблюдается обратная картина: в ряду ядро<оболочка<облако их доля растет, причем рост наиболее существенен для межгенных днРНК. Если в ядерной части их было около 10%, то в облаке они составляют более половины последовательностей. Для интронных днРНК тоже наблюдается аналогичное увеличение доли последовательностей, однако более умеренное по сравнению с межгенными днРНК. Таким образом, межгенные днРНК, по-видимому, представляют собой наиболее мобильную часть не кодирующего пан-транскриптома.

Сравнение скоростей замен показывает, что антисмысловые днРНК оказываются наиболее консервативными (рисунок 4.7Б). Гистограмма K_n для них смещена в сторону меньших значений, параметр α для гамма распределения наименьший из трех типов (0.37). Интересно, что наибольшее среднее значение

параметра α наблюдается для интронных днРНК (0.48), оно превышает значение для межгенных днРНК (0.43). Это позволяет предположить, что давление отбора меньше действует на интронные днРНК, в сравнении с остальными классами; вероятно, интронные днРНК являются более функционально нагруженными.

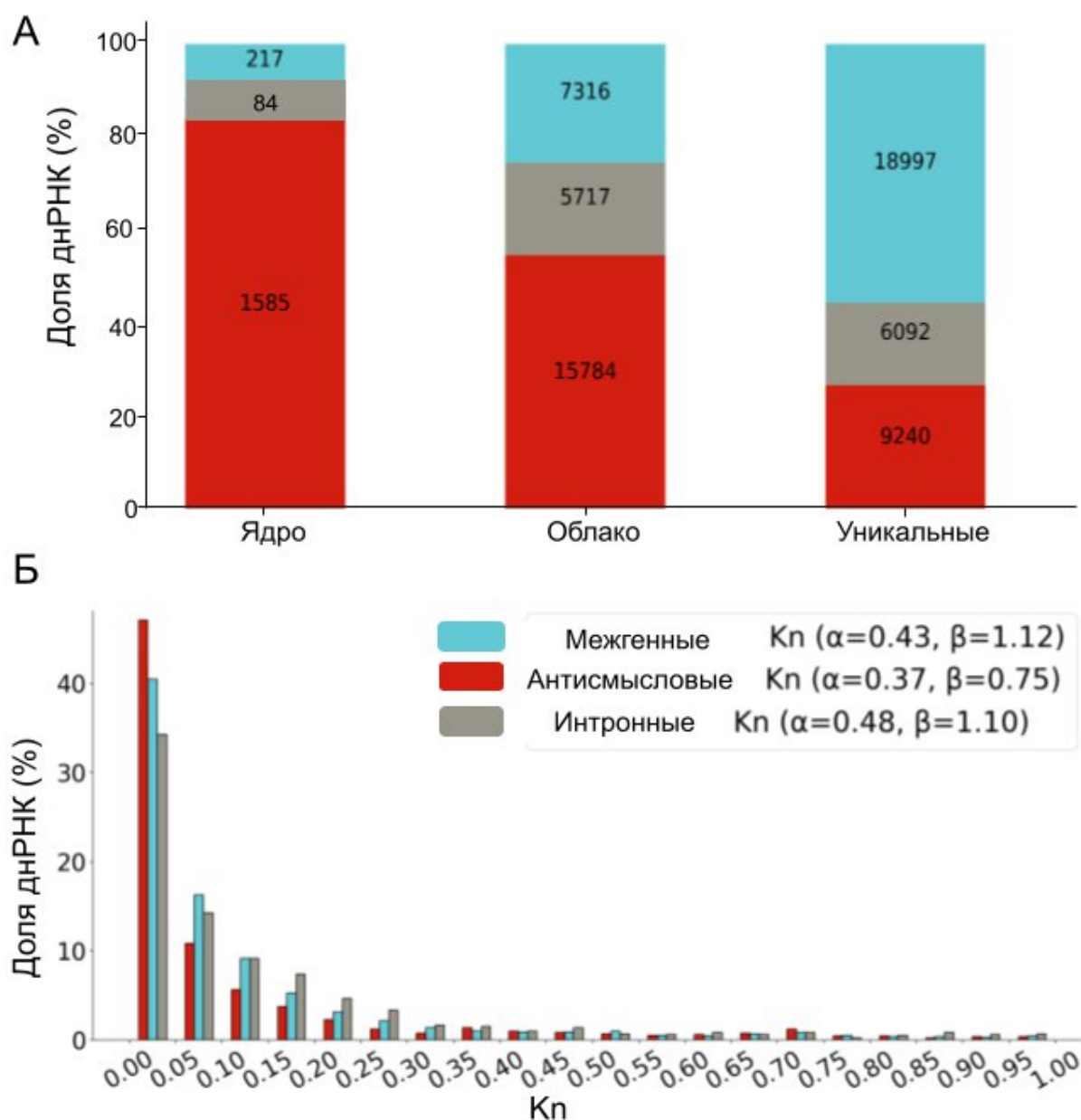


Рисунок 4.7. Анализ классов днРНК в пан-транскриптоме. А) Распределение классов днРНК относительно типов генов в пан-транскриптоме. По оси X представлены типы генов в пан-транскриптоме, по оси Y доля днРНК, которые принадлежат данному типу генов. Б) Распределение классов днРНК относительно их частоты нуклеотидных замен (K_n). По оси X представлены промежутки

значений K_n , по оси Y доля генов из каждого класса, соответствующих данному промежутку.

4.2 Обсуждение

4.2.1 Пан-транскриптом инбредных линий кукурузы в сравнении с другими пан-геномами и пан-транскриптомами

Анализ эволюции днРНК показал, что эти молекулы подвержены быстрому эволюционному изменению, которое приводит к тому, что только небольшая доля последовательностей разных видов оказывается консервативной. В настоящей работе сравнительный анализ проведен для днРНК кукурузы в рамках концепции пан-транскриптома, что позволяет оценить разнообразие их последовательностей, выявить закономерности изменчивости для большого количества последовательностей и сравнить их с особенностями изменчивости мРНК. При этом идентификация транскриптов как мРНК, так и днРНК проводилась на основе одних и тех же данных, одинаковым набором программ, что позволяет провести такое сравнение максимально корректно.

Мы выбрали для анализа данные, полученные в работе Hirsch et al. (2014), которые представляют пан-транскриптом 503 инбредных линий кукурузы. Однако глубина покрытия для этих библиотек относительно невелика (рисунок 4.8) и для большинства не превышает 20×. Также стоит отметить, что выделение РНК проводилось методом Trizol с использованием поли(А)-хвостов, что позволило минимизировать загрязнение рибосомальной РНК. Определённым ограничением такого подхода является то, что неполиаденилированные нкРНК остаются за пределами анализа. Однако во многих работах показано, что большинство днРНК подвергаются полиаденилированию (Mattick et al., 2023; Ziegler, 2017; Kornienko et al., 2024). Например, в работе по исследованию днРНК человека и мыши было показано, что более 70 % днРНК полиаденилированы (Ziegler, 2017). Более того, данные подходы выделения РНК часто используются в работах, посвященных исследованию днРНК (Kornienko et al., 2024; De Quattro et al., 2017; Han et al., 2019). Авторы аргументируют данный подход как

компромисс между качеством выделения и стоимостью исследования. Альтернативные методы, такие как GRO-seq и plaNET-seq, позволяют выявить полный спектр днРНК, но их высокая стоимость ограничивает применение для больших выборок образцов. Метод обеднения рРНК позволяет получить более полный набор днРНК в конкретном образце, однако он не всегда демонстрирует значимый прирост числа секвенированных днРНК по сравнению с поли(А)-селекцией и проигрывает ей в технологической простоте при массовом секвенировании (Dahlgren et al., 2020). Таким образом, метод, выбранный Hirsch et al., позволяет получить оптимальное количество днРНК за счет большого разнообразия образцов, что компенсирует потерю некоторого количества днРНК из-за метода выделения РНК.

Hirsch et al. (2014) использовали для сборки пан-транскриптома два подхода. В первом случае использовалось пулирование ридов из всех библиотек. Часть из них были картированы на референсный геном, оставшиеся использовались для *de novo* сборки (программы Velvet и Oases). Сборка *de novo* привела к реконструкции ~100 тысяч транскриптов, сформировавших ~30 тысяч наиболее достоверных транскриптов (RTAs). Из них ~22 тысячи обнаружили высокое сходство с белок кодирующими генами референсного генома, а ~8 тысяч определены как новые. Второй вариант сборки пан-транскриптома был основан на индивидуальной сборке 366 инбредных линий с наибольшим покрытием и затем объединении транскриптов в RTAs. Этот вариант сборки дал ~2 200 тысяч транскриптов, объединенных в ~750 тыс. RTAs. Эти данные были использованы для оценки типа пан-транскриптома (открытый или закрытый). В нашем конвейере мы использовали комбинированную сборку тремя современными программами транскриптов *de novo* для всех библиотек. Этот метод схож со вторым вариантом из работы Hirsch (2014). Для мРНК получено около 14 000 тысяч транскриптов и ~250 тысяч RTAs. Это количество оказывается между консервативной и расширенной оценками количества RTAs из работы Hirsch, а по порядку величин ближе ко второй. Однако точное сравнение результатов сборки транскриптома двух работ представляется вряд ли возможным,

вследствие разных подходов к сборке транскриптомов, разных порогов кластеризации и фильтрации.

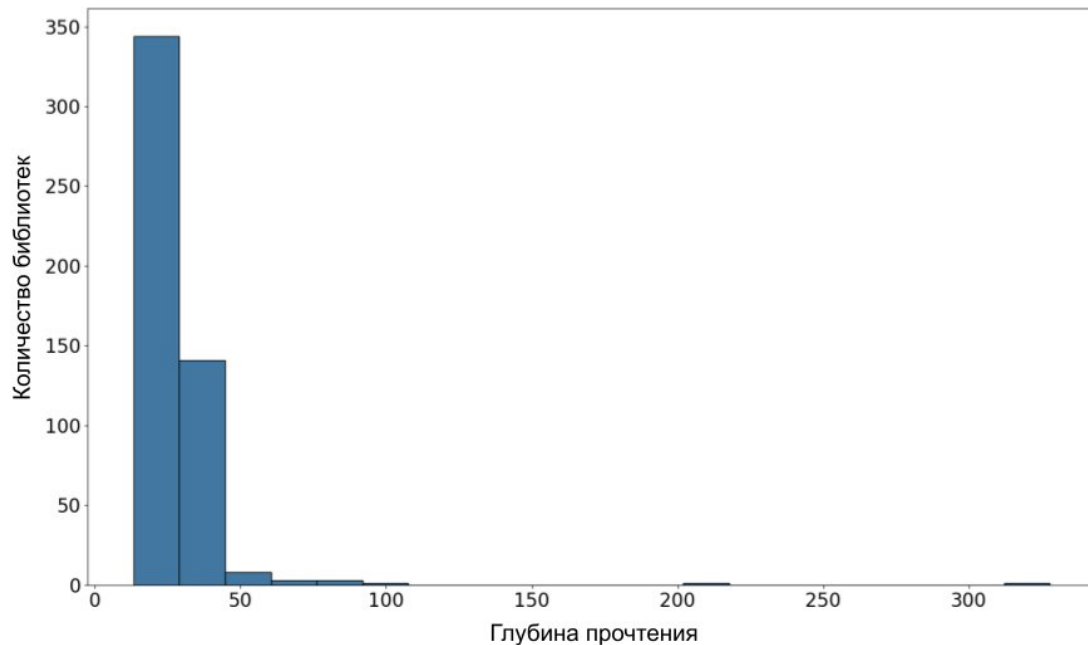


Рисунок 4.8. Распределение средней глубины прочтений в 503 библиотеках. Ось X средняя глубина прочтений, ось Y количество библиотек. Пик на глубине прочтения > 300 принадлежит библиотеке, полученной из референсного генома.

Как и в ранее проведенных работах (Sang et al., 2021, Zhang, 2014) мы показали, что состав днРНК является высоко вариабельным у инбредных линий кукурузы, в сравнении с мРНК. Об этом свидетельствует почти в три раза больший размер пан-транскриптома. Кроме того, для днРНК уникальная часть пан-транскриптома преобладает. Анализ количества транскриптов показал, что 95% пан-транскриптома мРНК представлено в 25 случайных образцах, а 99% в 93 образцах. Данные результаты близки к оценкам, полученным в работе (Gui et al. 2022), где насыщение до 95% генов в пан-геноме кукурузы достигалось на 27 образцах, а до 99% генов — на 93 образцах. Наши результаты показывают, что пан-транскриптом, кодирующий белки, у инбредных линий кукурузы имеет закрытый тип. Это совпадает с выводами Hirsch и др. (2014), а также с результатами исследования пан-генома кукурузы, выполненного на 721 особи, которое также подтверждает закрытый характер пан-генома (Gui et al., 2022).

Напротив, 95% транскриптов в пан-транскриптом днРНК представлены в 395 линиях, что в 15 раз превышает соответствующее значение для мРНК. Это дополнительно демонстрирует высокую изменчивость днРНК по сравнению с мРНК (рисунок 4.2А, раздел 4.1.3). Также стоит отметить, что насыщение пан-транскриптома днРНК происходит медленнее с увеличением числа инбредных линий, чем для мРНК. Эти результаты хорошо согласуются с данными, полученными в работе Kornienko et al. (Kornienko et al., 2024), в которой анализировали днРНК в транскриптомах 499 образцов *Arabidopsis thaliana*. Результаты показали, что в отличие от белок кодирующих генов, количество днРНК сильно зависит от размера выборки и не демонстрирует признаков насыщения даже при 460 образцах. Согласно оценкам степенного закона (Tettelin et al., 2008), пан-транскриптом днРНК кукурузы определяется как закрытый, подобно пан-транскриптому мРНК, несмотря на более низкую скорость насыщения (см. рисунок 4.2, раздел 4.1.3). Вопрос о типе пан-транскриптома днРНК важен для понимания возможных механизмов возникновения вариаций (ePAV). Kornienko et al. (Kornienko et al., 2024) показали, что изменчивость экспрессии днРНК у *Arabidopsis* в основном определяется эпигенетическими паттернами среди образцов. Это указывает на то, что общее число генов днРНК, закодированных в разных образцах, практически одинаково и предполагает закрытый характер пан-транскриптома днРНК.

Размер ядра пан-транскриптома мРНК составляет 8% от полного количества RTAs, что составляет ~20 тысяч транскриптов и сравнимо с количеством известных белок кодирующих генов в референсном геноме ~30 тысяч, (Hirsch et al., 2014). Исследования консервативности белок кодирующих генов в различных пан-геномах показали, что кластеры генов, входящих в ядро варьируют от 23% до 89%, кластеры генов входящих в оболочку — от 10% до 53%, а кластеры облака — от 0.1% до 43% (Jiang et al., 2025). Например, в *Brachypodium distachyon* было установлено, что в среднем отдельные линии в основном состояли из генов входящих в ядро пан-генома (73%), тогда как только 27% относились к оболочке или облаку (Gordon et al., 2017). С другой стороны, в пан-геноме ячменя количество уникальных генов было чуть больше, чем генов

в ядре (Guo et al., 2025). Тем не менее аналогично нашим данным, в большинстве случаев доля кластеров с уникальными генами оказывалась ниже, чем доля генов из ядра или оболочки, при этом их распределение варьируется между отдельными особями (Jiang et al., 2025).

В отличие от мРНК, пан-транскриптом днРНК показал преобладание последовательностей из уникальных генов (см. рисунок 4.4). Это подтверждает быструю эволюцию днРНК и преобладание среди них видоспецифичных последовательностей (Nitsche, 2017; Chen, 2022; Sang et al., 2021). Наши результаты также согласуются с исследованиями днРНК в популяции *A. thaliana* (Kornienko A. 2024), которые показали, что паттерны экспрессии днРНК сильно различаются между образцами: половина всех днРНК экспрессируется в одном образце и не экспрессируется в другом.

Мы наблюдали большие различия между нуклеотидными последовательностями для днРНК по сравнению с мРНК в рамках пан-транскриптома кукурузы, что хорошо согласуется с предыдущими исследованиями растений и животных (Chen и Zhu, 2022; Marques и Ponting, 2009; Necsulea et al., 2014; Deng et al, 2018). Оценки значений показателя Таджимы D в нашей работе оказались смещены в положительную сторону, как для мРНК, так и для днРНК. Для белок кодирующих генов наши результаты существенно отличаются от распределения показателя Таджимы D , полученного при анализе пан-генома 721 образца кукурузы (располагается в области отрицательных значений) (Gui et al., 2022). Однако следует иметь в виду, что, в отличие от работы (Gui et al., 2022), мы анализировали образцы инбредных линий кукурузы, для которых характерен дефицит аллелей низкой частоты по сравнению с ожидаемым уровнем, обусловленный возможным эффектом бутылочного горлышка, разделением популяции или балансирующим отбором (Ching et al., 2002). Отметим, что и этот тест демонстрирует отличие днРНК по сравнению с мРНК: распределение показателя Таджимы D у днРНК смещено в область больших значений.

Оценки параметров распределений несинонимических и синонимических замен для мРНК и замен в днРНК также показали большую изменчивость

последовательностей днРНК: для параметра α этих распределений характерно возрастание в ряду $K_a < K_s < K_n$, что соответствует увеличению скорости нуклеотидных замен. При этом скорость замен нуклеотидов K_n в последовательностях днРНК превышает даже скорости нейтральных замен K_s у мРНК. Интересно, что такой же тренд увеличения скорости замен мы наблюдаем для параметра α у трех типов замен в зависимости от принадлежности ядру, оболочке и облаку: $K_{\text{ядро}} < K_{\text{оболочка}} < K_{\text{облако}}$. Это характеризует консервативность последовательностей, принадлежащих разным частям пан-транскриптома в согласии с существующими представлениями об их эволюции (Gui et al., 2022; Golicz et al., 2016; Li et al., 2021).

В нашей работе также показано в согласии с ранее известными данными (Sang, 2022; Pronozin и Afonnikov, 2022), что по скорости замен из трех типов днРНК (антисмысловые, интронные и межгенные) по расположению в геноме наиболее консервативными являются антисмысловые, менее консервативны межгенные, а наиболее вариабельны интронные днРНК. Однако межгенные днРНК демонстрируют более высокое разнообразие eRAV: они наиболее часто представлены в облачной части пан-транскриптома. В пан-транскриптомах мРНК большинство уникальных генов относится к последовательностям с низкой экспрессией (Jiang et al, 2025; Gordon et al, 2017). Их функции обычно связаны с реакциями на окружающую среду и защитными механизмами, включая рецепторную и антиоксидантную активность, регуляцию генов и передачу сигналов (Golicz et al., 2020). Кроме того, они эволюционно моложе, чем гены из ядра (Golicz et al., 2020). Это позволяет предположить, что межгенные днРНК могут участвовать в регуляции стрессовых реакций растений и адаптации к окружающей среде. Примечательно, что несколько исследований показали связь межгенных днРНК с ответом растений на стресс (Palos et al., 2022; Golicz, 2018; Yin et al., 2021; Yu et al., 2025).

В целом наш анализ на основании изучения ряда характеристик вариабельности последовательностей РНК в транскриптомах кукурузы с помощью биоинформатических подходов к анализу мРНК и днРНК продемонстрировал закономерности их эволюции: высокое разнообразие состава

днРНК по сравнению с мРНК, высокая скорость замен у днРНК, более консервативная природа антисмысловых и межгенных днРНК по сравнению с интронными.

В ряде работ было ранее отмечено, что на анализ днРНК может неизбежно влиять целый ряд факторов: качество библиотек, качество сборки и аннотации генома (Chen и Zhu, 2022), выбор программ для идентификации и аннотации днРНК (Klapproth et al., 2021). Исследование методов реконструкции пан-генома на основе разных методов кластеризации последовательностей также показало, что выбор стратегии и метода объединения может влиять на оценку его структурных свойств (Manzano-Morales et al., 2023). В настоящей работе глубина секвенирования для данных RNA-seq довольно низкая. Это важное ограничение может приводить к недопредставленности транскриптов с низкой экспрессией в пан-транскриптоме. Низкая глубина секвенирования может повлиять на полноту наборов как мРНК, так и днРНК, однако этот фактор сильнее сказывается на некодирующих транскриптах из-за их более низкого уровня экспрессии (Chen, 2022). Второй причиной возможной недопредставленности обоих пан-транскриптомов является использование ткани побега на стадии первого листа. Известно, что экспрессия днРНК специфична для определенных тканей и стадий развития (Cabili et al., 2011; Liu et al., 2015). Поэтому часть транскриптов днРНК может быть пропущена при использовании одной или нескольких стадий развития, тканей или условий. Например, Kornienko et al. (Kornienko et al., 2024) продемонстрировали, что, несмотря на то, что количество локусов днРНК всегда увеличивалось с ростом числа образцов, количество использованных тканей оказывало еще большее влияние на полноту набора днРНК.

Помимо низкой глубины секвенирования и одной стадии развития в экспериментах RNA-seq, на оценку размера пан-транскриптома также может влиять порог экспрессии генов. В данной работе мы применили один и тот же порог для мРНК и днРНК ($TPM > 1$). Этот порог обычно принимается для удаления шума из набора мРНК (Wang, 2023; Wang, 2024; Herbert, 2021). Однако у большинства днРНК уровень экспрессии ниже, чем у мРНК (Chen, 2022). В связи с этим некоторые авторы применяют более низкие пороги для фильтрации

шума из набора днРНК, 0.3-0.5 TPM (Kornienko, 2024; Wang et al., 2020). Другие исследователи используют правило $TPM > 1$ для отбора днРНК (Guo et al., 2025; Guo et al., 2013; Gupta et al., 2010.), оправдывая этот выбор стремлением уменьшить более высокий уровень шума в данных днРНК. В нашем исследовании мы применили последний критерий. Это связано с низкой глубиной секвенирования, что может привести к увеличению шума. Порог сходства последовательностей, использованный при кластеризации мРНК и днРНК, является еще одним источником неопределенности в нашем анализе. Мы использовали разные значения для мРНК (80% идентичности) и днРНК (70% идентичности), подобранные эмпирически с учетом различной скорости замен аминокислот и нуклеотидов, рисунок 2.4, раздел 2.2.2. Таким образом несмотря на разницу в протоколах обработки данных, полученные результаты анализа последовательностей методически сопоставимы для мРНК и днРНК. Это обеспечивает объективную сравнительную оценку разнообразия днРНК и мРНК на основе данных пан-транскриптома.

Заключение

днРНК представляют собой класс линейных или кольцевых молекул РНК длиной от 200 нуклеотидов, некодирующих белок. Долгое время их функции были слабо изучены, и эти молекулы рассматривались как транскрипционный шум. Однако современные исследования показали, что у растений днРНК участвуют в регуляции устойчивости к холодовому, солевому и тепловому стрессу, влияют на устойчивость к гипоксии, принимают участие в развитии плодов, корней и листьев и в большом количестве других процессов роста и развития растений, ответа на стресс. Однако, несмотря на рост количества исследований, связанных с данными молекулами, они по-прежнему остаются малоизученными. Это связано со структурным сходством с мРНК, низким уровнем экспрессии и быстрыми темпами эволюции. Сравнительный анализ последовательностей днРНК между различными видами растений показал, что большая часть последовательностей представлена только у отдельных видов, т.е. являются внутривидовыми. Поливидовые последовательности, представленные

у нескольких видов, встречаются реже. При этом доля поливидовых днРНК резко падает по мере увеличения эволюционного расстояния между видами.

Таким образом, высокая вариабельность и преобладание внутривидовых днРНК не позволяют в полной мере использовать их межвидовой сравнительный анализ для выявления функций. Однако технологии секвенирования в настоящее время позволяют получать для вида не только единственную референсную последовательность генома, но и ДНК других его представителей. Это позволяет изучать структуры генов, отсутствующих в референсной последовательности. Разнообразие генов у представителей вида может быть описано моделью пан-генома или пан-транскриптома.

Данная работа посвящена поиску и анализу днРНК в пан-транскриптоме кукурузы. На первом этапе мы разработали биоинформатический конвейер ICAnnoLncRNA, который позволяет проводить предсказание и аннотацию днРНК растений. Преимущество данного конвейера в сравнении с аналогами заключается в использовании методов машинного обучения и дополнительных параметров фильтрации, которые можно подобрать индивидуально для специфического референсного генома. Конвейер был апробирован на транскриптомных библиотеках кукурузы и *A. thaliana*. Полученные результаты согласуются как между исследуемыми видами (кукурузой и арабидопсисом), так и с данными литературы для других растений. Это свидетельствует не только о высоком качестве предсказания днРНК с помощью разработанного конвейера, но и о достоверности аннотации, получаемой исследователями.

На втором этапе мы разработали биоинформатический конвейер РТА. Он позволяет осуществлять автоматизированную сборку, кластеризацию и сравнительный анализ кодирующих и не кодирующих частей пан-транскриптома. Разработанный конвейер был применен для анализа 503 разнообразных инбредных линий кукурузы. Было показано, что оба пан-транскриптома (кодирующий и не кодирующий) являются закрытыми. Однако, несмотря на идентичный тип пан-транскриптомов, состав днРНК является более вариабельным у инбредных линий кукурузы по сравнению с мРНК. Об этом свидетельствует почти в три раза больший размер пан-транскриптома. Также это

подтверждается тем, что пан-транскриптом днРНК в основном состоит из группы транскриптов, принадлежащих облаку, а белок кодирующих — в основном из группы оболочки и ядра. Полученные результаты согласуются с литературными данными, согласно которым днРНК подвержены быстрому эволюционному отбору и являются более видоспецифичными последовательностями по сравнению с мРНК.

Более высокая вариабельность также подтверждается распределением несинонимических и синонимических замен для мРНК и днРНК: для параметра α этих распределений характерно возрастание в ряду $K_a < K_s < K_n$, что соответствует увеличению скорости нуклеотидных замен. При этом скорости замен нуклеотидов K_n в последовательностях днРНК превышают даже скорости нейтральных замен K_s у мРНК. Это может говорить об отличии режима эволюции для днРНК по сравнению с мРНК. Более высокий показатель K_n над K_s может свидетельствовать о том, что режим эволюции днРНК отличен от нейтрального, что способствует их быстрой дивергенции и возникновению новых видоспецифичных регуляторных элементов.

Проведенный более детальный анализ трех классов показал, что по скорости замен наиболее консервативными являются антисмысловые, менее консервативны межгенные, а наиболее вариабельны интронные днРНК. Учитывая, что в пан-транскриптомах мРНК большинство уникальных генов относится к последовательностям с низкой экспрессией, а их функции обычно связаны с реакциями на окружающую среду и защитными механизмами, а также то, что они эволюционно моложе генов из ядра, можно предположить, что межгенные и интронные днРНК могут участвовать в регуляции стрессовых реакций растений и адаптации к окружающей среде.

Выводы

1. Разработан биоинформатический конвейер ICApnoLncRNA для идентификации в транскриптомных данных растений длинных некодирующих РНК и их комплексного анализа, включая определение типа локализации в геноме, определение экзон-интронной структуры генов, поиск гомологов в базах данных последовательностей днРНК растений, анализ экспрессии в различных тканях и органах растений.
2. Разработан биоинформатический конвейер РТА для сборки и аннотации кодирующих и длинных некодирующих РНК в транскриптомах растений, который позволяет анализировать их разнообразие, особенности структуры и эволюции в рамках модели пан-генома.
3. Впервые на основе сравнения структурных характеристик пан-транскриптомов мРНК и днРНК у растений инбредных линий кукурузы показан их закрытый характер, пан-транскриптом мРНК преимущественно сформирован эволюционно древними генами «ядра» и «оболочки», тогда как основу пан-транскриптома днРНК составляют молодые и переменные последовательности «облака».
4. Показано, что скорость нуклеотидных замен в днРНК (K_n) выше, чем скорости синонимичных (K_s) и несинонимичных (K_a) замен в мРНК, причем данная закономерность характерна для ядра, оболочки и облака пан-транскриптома.
5. Проведённый впервые для растений анализ структурных и эволюционных характеристик пан-транскриптома днРНК инбредных линий кукурузы показал, что антисмысловые днРНК являются наиболее консервативными как по большей представленности в ядре, так и по темпам нуклеотидных замен, интронные днРНК характеризуются наибольшей скоростью замен; межгенные днРНК преобладают среди облака, что свидетельствует об их вовлеченности в процессы адаптации растений к стрессу и адаптации к условиям среды.

Апробация результатов

Тезисы:

1. **Pronozin A.**, Afonnikov D. Identification and structural analysis of long noncoding RNAs - MCCMB 2021
2. **Pronozin A.**, Afonnikov D. Prediction, classification and annotation of long noncoding RNAs - CBB6
3. **Pronozin A.**, Afonnikov D. Identification and classification of long noncoding RNAs - Марчуковские научные чтения - 2021.
4. **Pronozin A.**, Afonnikov D. Identification and structural features analysis of long noncoding RNAs - Plant Genetics, Genomics, Bioinformatics, and Biotechnology. -2021
5. **Pronozin A.**, Afonnikov D. ICAnnoLncRNA : automatic pipeline of identification, Classification and annotation of long noncoding RNAs. - Bioinformatics of Genome Regulation and Structure/Systems Biology”, 2022
6. **Pronozin A.**, Afonnikov D.The ICAnnoLncRNA pipeline for a Long-Non-coding-RNA identification and Annotation in Transcriptomic Data - PlantGen2023
7. **Пронозин А. Ю.**, Афонников Д.А. Поиск и анализ длинных некодирующих РНК в масштабах пан-транскриптома кукурузы. - VIII Съезд Вавиловского общества генетиков и селекционеров 2024
8. **Pronozin A.**, Afonnikov D.Pan-transcriptome-based long non-coding RNAs prediction and analysis // Bioinformatics of Genome Regulation and Structure/Systems Biology”, 2024 - doi 10.18699/bgrs2024-1.4-13
9. **Pronozin A.**, Afonnikov D., Development of a bioinformatics pipeline for GBS data processing // Bioinformatics of Genome Regulation и Structure/Systems Biology”, 2024 - doi 10.18699/bgrs2024-1.4-13
10. **Pronozin A.**, Afonnikov D. A pan-transcriptome scale study of long non-coding RNA diversity in maize —PlantGen2025

Публикации:

1. Pronozin AY, Bragina MK, Salina EA. Crop pangenomes. *Vavilovskii Zhurnal Genet Selektsii*. 2021 Feb;25(1):57-63. doi: 10.18699/VJ21.007. PMID: 34901703; PMCID: PMC8629360.
2. Pronozin AY, Afonnikov DA. ICAnnoLncRNA pipeline for lncRNA search и annotation in transcriptomic data: application for maize. *Genes* 2023
3. Pronozin, A.Y., Afonnikov, D.A. The Role of Long Noncoding RNAs in Plants. *Russ J Genet* 61, 1-18 (2025). <https://doi.org/10.1134/S1022795424701345>
4. Pronozin AY, Karetnikov DI, Shmakov NA, Bocharnikova ME, Afonnikova SD, Afonnikov DA, Kolchanov NA. CropGene: a software package for the analysis of genomic and transcriptomic data of agricultural plants. *Vavilovskii Zhurnal Genet Selektsii*. 2025 Apr;29(2):320-329. doi: 10.18699/vjgb-25-35. PMID: 40264806; PMCID: PMC12011622.
5. Pronozin AY, Shmakov NA, Afonnikov DA Diversity of lncRNAs in the pan-transcriptome of maize inbred lines.

Список литературы

1. Abendroth L. et al. Corn Growth and Development. - 2011.
2. Achawanantakun R. et al. LncRNA-ID: Long non-coding RNA IDentification using balanced random forests // Bioinformatics. - 2015. - Т. 31. - №. 24. - С. 3897-3905.
3. Ahmad P. et al. Long non-coding RNAs and their potential roles in the vector-host-pathogen triad // Life. - 2021. - Т. 11. - №. 1. - С. 56.
4. Akhunov E. D. et al. Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes // BMC Genomics. - 2010. - Т. 11. - №. 1. - С. 702.
5. Alcaraz L. D. et al. Understanding the evolutionary relationships and major traits of Bacillus through comparative genomics // BMC Genomics. - 2010. - Т. 11. - №. 1. - С. 332.
6. Amaral P. P. et al. Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci // Genome Biology. - 2018. - Т. 19. - №. 1. - С. 32.
7. Amaral P. P., Mattick J. S. Noncoding RNA in development // Mammalian Genome. - 2008. - Т. 19. - №. 7-8. - С. 454-492.
8. Amor B. B. et al. Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses // Genome research. - 2009. - Т. 19. - №. 1. - С. 57-69.
9. Ariel F. et al. Battles and hijacks: noncoding transcription in plants // Trends in plant science. - 2015. - Т. 20. - №. 6. - С. 362-371.
10. Au P. C. K., Zhu Q.-H. Identification of lncRNAs Using Computational and Experimental Approaches // Regulatory RNAs / Под ред. B. Mallick, Z. Ghosh. - Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. - С. 319-340.
11. Augoff K. et al. miR-31 and its host gene lncRNA LOC554202 are regulated by promoter hypermethylation in triple-negative breast cancer // Molecular Cancer. - 2012. - Т. 11. - №. 1. - С. 5.

12. Azlan A. et al. Systematic Identification and Characterization of *Aedes Aegypti* Long Noncoding RNAs (LncRNAs) // *Scientific Reports*. - 2019. - T. 9. - №. 1. - C. 12147.
13. Babak T., Blencowe B. J., Hughes T. R. A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription // *BMC Genomics*. - 2005. - T. 6. - №. 1. - C. 104.
14. Baek J. et al. LncRNAnet: long non-coding RNA identification using deep learning // *Bioinformatics*. - 2018. - T. 34. - №. 22. - C. 3889-3897.
15. Bai Y. et al. PlncRNADB: A Repository of Plant lncRNAs and lncRNA-RBP Protein Interactions // *Current Bioinformatics*. - 2019. - T. 14. - №. 7. - C. 621-627.
16. Bardou F. et al. Dual RNAs in plants // *Biochimie*. - 2011. - T. 93. - №. 11. - C. 1950-1954.
17. Bardou F. et al. Long noncoding RNA modulates alternative splicing regulators in *Arabidopsis* // *Developmental cell*. - 2014. - T. 30. - №. 2. - C. 166-176.
18. Beadle G. W. *Teosinte and the origin of maize*. - 1939.
19. Beltran M. et al. A natural antisense transcript regulates *Zeb2/Sip1* gene expression during *Snail1*-induced epithelial-mesenchymal transition // *Genes & development*. - 2008. - T. 22. - №. 6. - C. 756-769.
20. Bernard D. et al. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression // *The EMBO Journal*. - 2010. - T. 29. - №. 18. - C. 3082-3093.
21. Bhatia G. et al. Present Scenario of Long Non-Coding RNAs in Plants // *Non-coding RNA*. - 2017. - T. 3. - №. 2. - C. 16.
22. Blythe A. J., Fox A. H., Bond C. S. The ins and outs of lncRNA structure: How, why and what comes next? // *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*. - 2016. - T. 1859. - №. 1. - C. 46-58.
23. Bolduc N. et al. Unraveling the *KNOTTED1* Regulatory Network in Maize Meristems // *Genes & Development*. - 2012. - T. 26. - №. 15. - C. 1685-1690.

24. Bolser D. et al. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data // Plant Bioinformatics. - Springer, 2016. - C. 115-140.
25. Boratyn G. M. et al. BLAST: a more efficient report with usability improvements // Nucleic acids research. - 2013. - T. 41. - №. W1. - C. W29-W33.
26. Bray N. L. et al. Near-Optimal Probabilistic RNA-Seq Quantification // Nature Biotechnology. - 2016. - T. 34. - №. 5. - C. 525-527.
27. Britto-Kido S. D. A. et al. Natural Antisense Transcripts in Plants: A Review and Identification in Soybean Infected with *Phakopsora pachyrhizi* SuperSAGE Library // The Scientific World Journal. - 2013. - T. 2013. - №. 1. - C. 219798.
28. Bryzghalov O., Makałowska I., Szcześniak M. W. lncEvo: automated identification and conservation study of long noncoding RNAs // BMC Bioinformatics. - 2021. - T. 22. - №. 1. - C. 59.
29. Buckler E. S., Stevens N. M. Maize origins, domestication, and selection // Darwin's harvest: New approaches to the origins, evolution, and conservation of crops. - Columbia University Press, 2006. - C. 67-90.
30. Budak H., Kaya S. B., Cagirici H. B. Long non-coding RNA in plants in the era of reference sequences // Frontiers in Plant Science. - 2020. - T. 11. - C. 276.
31. Bushmanova E., Antipov D., Lapidus A., Suvorov V., Prjibelski A. D. rnaQUAST: a quality assessment tool for de novo transcriptome assemblies // Bioinformatics. - 2016. - T. 32. - №. 14. - C. 2210-2212.
32. Butler J. et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads // Genome research. - 2008. - T. 18. - №. 5. - C. 810-820.
33. Cabili M. N. et al. Integrative Annotation of Human Large Intergenic Noncoding RNAs Reveals Global Properties and Specific Subclasses // Genes & Development. - 2011. - T. 25. - №. 18. - C. 1915-1927.
34. Cabili M. N. et al. Localization and Abundance Analysis of Human LncRNAs at Single-Cell and Single-Molecule Resolution // Genome Biology. - 2015. - T. 16. - №. 1. - C. 20.
35. Cai X., Cullen B. R. The imprinted H19 noncoding RNA is a primary microRNA precursor // RNA. - 2007. - T. 13. - №. 3. - C. 313-316.

36. Cairns J. E., Prasanna B. M. Developing and deploying climate-resilient maize varieties in the developing world // *Current Opinion in Plant Biology*. - 2018. - T. 45. - C. 226-230.
37. Calixto C. P. et al. Cold-dependent expression and alternative splicing of *Arabidopsis* long non-coding RNAs // *Frontiers in Plant Science*. - 2019. - T. 10. - C. 235.
38. Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. (2008) "BLAST+: architecture and applications." *BMC Bioinformatics* 10:421. PubMed
39. Campalans A., Kondorosi A., Crespi M. Enod40, a short open reading frame-containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in *Medicago truncatula* // *The Plant Cell*. - 2004. - T. 16. - №. 4. - C. 1047-1059.
40. Carlson C. S. et al. Genomic regions exhibiting positive selection identified from dense genotype data // *Genome research*. - 2005. - T. 15. - №. 11. - C. 1553-1565.
41. Ceriani L., Verme P. The origins of the Gini index: extracts from *Variabilità e Mutabilità* (1912) by Corrado Gini // *The Journal of Economic Inequality*. - 2012. - T. 10. - №. 3. - C. 421-443.
42. Cerveau N., Jackson D. J. Combining independent de novo assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms // *BMC Bioinformatics*. - 2016. - T. 17. - №. 1. - C. 525.
43. Chaudhari N. M., Gupta V. K., Dutta C. BPGA-an ultra-fast pan-genome analysis pipeline // *Scientific reports*. - 2016. - T. 6. - №. 1. - C. 24373.
44. Chekanova J. A. et al. Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the *Arabidopsis* transcriptome // *Cell*. - 2007. - T. 131. - №. 7. - C. 1340-1353.
45. Chen J., Wang H., Yao Y. Experimental study of nonlinear ultrasonic behavior of soil materials during the compaction // *Ultrasonics*. - 2016. - T. 69. - C. 19-24.

46. Chen J., Zhong Y., Qi X. LncRNA TCONS_00021861 is functionally associated with drought tolerance in rice (*Oryza sativa* L.) via competing endogenous RNA regulation // *BMC Plant Biology*. - 2021. - T. 21. - №. 1. - C. 410.
47. Chen K., Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs // *Nature Reviews Genetics*. - 2007. - T. 8. - №. 2. - C. 93-103.
48. Chen L., Zhu Q.-H. The evolutionary landscape and expression pattern of plant lincRNAs // *RNA Biology*. - 2022. - T. 19. - №. 1. - C. 1190-1207.
49. Chen L., Zhu Q.-H., Kaufmann K. Long non-coding RNAs in plants: emerging modulators of gene activity in development and stress responses // *Planta*. - 2020. - T. 252. - №. 5. - C. 92.
50. Chen S. et al. fastp: An Ultra-Fast All-in-One FASTQ Preprocessor // *Bioinformatics*. - 2018. - T. 34. - №. 17. - C. i884-i890.
51. Cheng F. et al. Syntenic gene analysis between *Brassica rapa* and other Brassicaceae species // *Frontiers in plant science*. - 2012. - T. 3. - C. 198.
52. Cheng W. et al. The Putative Smallest Introns in the *Arabidopsis* Genome // *Genome Biology and Evolution*. - 2018. - T. 10. - №. 10. - C. 2551-2557.
53. Cheung F. et al. Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology // *BMC Genomics*. - 2006. - T. 7. - №. 1. - C. 272.
54. Chillón I. et al. Native purification and analysis of long RNAs // *Methods in enzymology*. - Elsevier, 2015. - C. 3-37.
55. Ching A. et al. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines // *BMC Genetics*. - 2002. - T. 3. - №. 1. - C. 19.
56. Chowdhary A., Satagopam V., Schneider R. Long non-coding RNAs: mechanisms, experimental, and computational approaches in identification, characterization, and their biomarker potential in cancer // *Frontiers in genetics*. - 2021. - T. 12. - C. 649619.
57. Chung P. J. et al. Transcriptome profiling of drought responsive noncoding RNAs and their target genes in rice // *BMC Genomics*. - 2016. - T. 17. - №. 1. - C. 563.

58. Churchman L. S., Weissman J. S. Native elongating transcript sequencing (NET-seq) // *Current Protocols in Molecular Biology*. - 2012. - T. 98. - №. 1. - C. 14.4.1-14.4.17.
59. Conn V. M. et al. A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation // *Nature plants*. - 2017. - T. 3. - №. 5. - C. 1-5. da Costa Negri T., Paschoal A. R., Alves W. A. L. Comparison tools for lncRNA identification: Analysis among plants and humans // *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. - IEEE, 2020. - C. 1-8.
60. Dahlgren A. R., Eivers S. S., Geor R. J., Spencer T. E. Comparison of poly-A⁺ selection and rRNA depletion in detection of lncRNA in two equine tissues using RNA-seq // *Non-coding RNA*. - 2020. - T. 6. - №. 3. - C. 32.
61. Danecek P. et al. The variant call format and VCFtools // *Bioinformatics*. - 2011. - T. 27. - №. 15. - C. 2156-2158.
62. Danecek P. et al. Twelve years of SAMtools and BCFtools // *Gigascience*. - 2021. - T. 10. - №. 2. - C. giab008.
63. De Quattro C., Pè M. E., Bertolini E. Long noncoding RNAs in the model species *Brachypodium distachyon* // *Scientific reports*. - 2017. - T. 7. - №. 1. - C. 11252.
64. Deng P. et al. Conservation analysis of long non-coding RNAs in plants // *Science China Life Sciences*. - 2018. - T. 61. - №. 2. - C. 190-198.
65. Derrien T. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression // *Genome research*. - 2012. - T. 22. - №. 9. - C. 1775-1789.
66. Deveson I. W., Hardwick S. A., Mercer T. R., Mattick J. S. The dimensions, dynamics, and relevance of the mammalian noncoding transcriptome // *Trends in Genetics*. - 2017. - T. 33. - №. 7. - C. 464-478.
67. Di Tommaso P. et al. Nextflow enables reproducible computational workflows // *Nature biotechnology*. - 2017. - T. 35. - №. 4. - C. 316-319.
68. Ding J. et al. A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice // *Proceedings of the National Academy of Sciences*. - 2012. - T. 109. - №. 7. - C. 2654-2659.

69. Ding W., Baumdicker F., Neher R. A. panX: pan-genome analysis and exploration // *Nucleic acids research*. - 2018. - T. 46. - №. 1. - C. e5.
70. Dinger M. E. et al. Differentiating protein-coding and noncoding RNA: challenges and ambiguities // *PLoS computational biology*. - 2008. - T. 4. - №. 11. - C. e1000176.
71. Doebley J. F., Goodman M., Stuber C. W. Isoenzymatic variation in *Zea* (gramineae) // *Systematic Botany*. - 1984. - C. 203-218.
72. Doebley J. Molecular systematics of *Zea* (Gramineae). - 1990.
73. Dolgin E. Maize genome mapped // *Nature News*. - 2009. - C. 1098.
74. Drolet M., Bi X., Liu L. F. Hypernegative supercoiling of the DNA template during transcription elongation in vitro // *Journal of Biological Chemistry*. - 1994. - T. 269. - №. 3. - C. 2068-2074.
75. Du Q. et al. The PILNCR1-miR399 regulatory module is important for low phosphate tolerance in maize // *Plant Physiology*. - 2018. - T. 177. - №. 4. - C. 1743-1753.
76. Duret L. et al. The Xist RNA Gene Evolved in Eutherians by Pseudogenization of a Protein-Coding Gene // *Science*. - 2006. - T. 312. - №. 5780. - C. 1653-1655.
77. Edgar R. C. Search and Clustering Orders of Magnitude Faster than BLAST // *Bioinformatics*. - 2010. - T. 26. - №. 19. - C. 2460-2461.
78. Evangelistella C. et al. De novo assembly, functional annotation, and analysis of the giant reed (*Arundo donax* L.) leaf transcriptome provide tools for the development of a biofuel feedstock // *Biotechnology for Biofuels*. - 2017. - T. 10. - №. 1. - C. 138.
79. Everaert C. et al. Benchmarking of RNA-Sequencing Analysis Workflows Using Whole-Transcriptome RT-QPCR Expression Data // *Scientific Reports*. - 2017. - T. 7. - №. 1. - C. 1559.
80. Faghihi M. A. et al. Evidence for natural antisense transcript-mediated inhibition of microRNA function // *Genome Biology*. - 2010. - T. 11. - №. 5. - C. R56.
81. Fatica A., Bozzoni I. Long Non-Coding RNAs: New Players in Cell Differentiation and Development // *Nature Reviews Genetics*. - 2014. - T. 15. - №. 1. - C. 7-21.

82. Federhen S. The NCBI taxonomy database // Nucleic acids research. - 2012. - T. 40. - №. D1. - C. D136-D143.
83. Fickett J. W. Recognition of protein coding regions in DNA sequences // Nucleic acids research. - 1982. - T. 10. - №. 17. - C. 5303-5318.
84. Fickett J. W., Tung C.-S. Assessment of protein coding measures // Nucleic acids research. - 1992. - T. 20. - №. 24. - C. 6441-6450.
85. Flynn R. A., Chang H. Y. Active chromatin and noncoding RNAs: an intimate relationship // Current opinion in genetics & development. - 2012. - T. 22. - №. 2. - C. 172-178.
86. Franco-Zorrilla J. M. et al. Target mimicry provides a new mechanism for regulation of microRNA activity // Nature genetics. - 2007. - T. 39. - №. 8. - C. 1033-1037.
87. Frankish A. et al. GENCODE 2021 // Nucleic acids research. - 2021. - T. 49. - №. D1. - C. D916-D923.
88. Gao L. et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor // Nature genetics. - 2019. - T. 51. - №. 6. - C. 1044-1051.
89. Geniza M., Jaiswal P. Tools for Building de Novo Transcriptome Assembly // Current Plant Biology. - 2017. - T. 11. - C. 41-45.
90. Gilbert D. G. Genes of the pig, *Sus scrofa*, reconstructed with EvidentialGene // PeerJ. - 2019. - T. 7. - C. e6374.
91. Golicz A. A. et al. Pangenomics comes of age: from bacteria to plant and animal applications // Trends in Genetics. - 2020. - T. 36. - №. 2. - C. 132-145.
92. Golicz A. A. et al. The pangenome of an agronomically important crop plant *Brassica oleracea* // Nature communications. - 2016. - T. 7. - №. 1. - C. 13390.
93. Golicz A. A., Singh M. B., Bhalla P. L. The Long Intergenic Noncoding RNA (LincRNA) Landscape of the Soybean Genome // Plant Physiology. - 2018. - T. 176. - №. 3. - C. 2133-2147.
94. Gong Y. et al. lncRNA-screen: an interactive platform for computationally screening long non-coding RNAs in large genomics datasets // BMC Genomics. - 2017. - T. 18. - №. 1. - C. 434.

95. Gordon S. P. et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure // *Nature Communications*. - 2017. - T. 8. - №. 1. - C. 2184.
96. Grabherr M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome // *Nature biotechnology*. - 2011. - T. 29. - №. 7. - C. 644-652.
97. Graf J., Kretz M. From structure to function: Route to understanding lncRNA mechanism // *BioEssays*. - 2020. - T. 42. - №. 12. - C. 2000027.
98. Griffiths-Jones S. miRBase: The MicroRNA Sequence Database // *MicroRNA Protocols* / Под ред. Y. Shao-Yao. - New Jersey: Humana Press, 2006. - C. 129-138.
99. Gui S. et al. A pan-Zea genome map for enhancing maize improvement // *Genome Biology*. - 2022. - T. 23. - №. 1. - C. 178.
100. Guo W. et al. A barley pan-transcriptome reveals layers of genotype-dependent transcriptional complexity // *Nature Genetics*. - 2025. - T. 57. - №. 2. - C. 441-450.
101. Guo X. et al. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks // *Nucleic acids research*. - 2013. - T. 41. - №. 2. - C. e35.
102. Gupta R. A. et al. Long Non-Coding RNA HOTAIR Reprograms Chromatin State to Promote Cancer Metastasis // *Nature*. - 2010. - T. 464. - №. 7297. - C. 1071-1076.
103. Guttman M. et al. Ab Initio Reconstruction of Cell Type-Specific Transcriptomes in Mouse Reveals the Conserved Multi-Exonic Structure of LincRNAs // *Nature Biotechnology*. - 2010. - T. 28. - №. 5. - C. 503-510.
104. Hou M., Pang S. Plant pan-genomics: Opportunities, advances, and challenges // *Journal of Data Science and Intelligent Systems*. - 2024.
105. Hölzer M., Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers // *Gigascience*. - 2019. - T. 8. - №. 5. - C. giz039.

106. Hübner S. et al. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance // *Nature plants*. - 2019. - T. 5. - №. 1. - C. 54-62.
107. Haas B. J. et al. De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis // *Nature Protocols*. - 2013. - T. 8. - №. 8. - C. 1494-1512.
108. Halstead M. M. et al. Large-Scale Multiplexing Permits Full-Length Transcriptome Annotation of 32 Bovine Tissues from a Single Nanopore Flow Cell // *Frontiers in Genetics*. - 2021. - T. 12. - C. 664260.
109. Han L. et al. New lncRNA annotation reveals extensive functional divergence of the transcriptome in maize // *Journal of Integrative Plant Biology*. - 2019. - T. 61. - №. 4. - C. 394-405.
110. Han S. et al. LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property // *Briefings in bioinformatics*. - 2019. - T. 20. - №. 6. - C. 2009-2027.
111. Hansey C. N., Johnson J. M., Sekhon R. S., Kaeppler S. M., de Leon N. Genetic diversity of a maize association population with restricted phenology // *Crop Science*. - 2010. - T. 51. - №. 2. - C. 704-715.
112. Hao Z. et al. Genome-wide identification, characterization and evolutionary analysis of long intergenic noncoding RNAs in cucumber // *PloS one*. - 2015. - T. 10. - №. 3. - C. e0121800.
113. Hawkes E. J. et al. COOLAIR antisense RNAs form evolutionarily conserved elaborate secondary structures // *Cell reports*. - 2016. - T. 16. - №. 12. - C. 3087-3096.
114. Herbert D. B. et al. Transcriptome analysis reveals major transcriptional changes during regrowth after mowing of red clover (*Trifolium pratense*) // *BMC Plant Biology*. - 2021. - T. 21. - №. 1. - C. 95.
115. Hetzel J., Duttke S. H., Benner C., Chory J. Nascent RNA sequencing reveals distinct features in plant transcription // *Proceedings of the National Academy of Sciences*. - 2016. - T. 113. - №. 43. - C. 12316-12321.

116. Hezroni H. et al. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species // *Cell reports*. - 2015. - T. 11. - №. 7. - C. 1110-1122.
117. Hirsch C. N. et al. Insights into the maize pan-genome and pan-transcriptome // *The Plant Cell*. - 2014. - T. 26. - №. 1. - C. 121-135.
118. Hirsch J. et al. Characterization of 43 non-protein-coding mRNA genes in *Arabidopsis*, including the MIR162a-derived transcripts // *Plant physiology*. - 2006. - T. 140. - №. 4. - C. 1192-1204.
119. Howe K. L. et al. Ensembl Genomes 2020—Enabling Non-Vertebrate Genomic Research // *Nucleic Acids Research*. - 2020. - T. 48. - №. D1. - C. D689-D695.
120. Huang L. et al. Systematic identification of long non-coding RNAs during pollen development and fertilization in *Brassica rapa* // *The Plant Journal*. - 2018. - T. 96. - №. 1. - C. 203-222.
121. Huang X., Li S. Z., Wang Y. Jensen-shannon boosting learning for object recognition // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). - IEEE, 2005. - C. 144-149.
122. Huarte M., Rinn J. L. Large non-coding RNAs: missing links in cancer? // *Human molecular genetics*. - 2010. - T. 19. - №. R2. - C. R152-R161.
123. Huffines A. K., Schneider D. A. Differential impact of divalent metals on native elongating transcript sequencing (NET-seq) protocols for RNA polymerases I and II // *PLOS ONE*. - 2025. - T. 20. - №. 2. - C. e0315595.
124. Hufford M. B., Seetharam A. S., Woodhouse M. R., Chougule K. M., Ou S., Liu J., et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes // *Science*. - 2021. - T. 373. - №. 6555. - C. 655-662.
125. Hung T. et al. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters // *Nature genetics*. - 2011. - T. 43. - №. 7. - C. 621-629.
126. Huntzinger E., Izaurralde E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay // *Nature reviews genetics*. - 2011. - T. 12. - №. 2. - C. 99-110.

127. Hurgobin B., Edwards D. SNP discovery using a pangenome: has the single reference approach become obsolete? // *Biology*. - 2017. - T. 6. - №. 1. - C. 21.
128. Hutchinson J. N. et al. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains // *BMC Genomics*. - 2007. - T. 8. - №. 1. - C. 39.
129. Ilik I. A. et al. Tandem Stem-Loops in RoX RNAs Act Together to Mediate X Chromosome Dosage Compensation in *Drosophila* // *Molecular cell*. - 2013. - T. 51. - №. 2. - C. 156-173.
130. Iyer M. K. et al. The Landscape of Long Noncoding RNAs in the Human Transcriptome // *Nature Genetics*. - 2015. - T. 47. - №. 3. - C. 199-208.
131. Jain P., Krishnan N. M., Panda B. Augmenting transcriptome assembly by combining de novo and genome-guided tools // *PeerJ*. - 2013. - T. 1. - C. e133.
132. Jalili V. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update // *Nucleic acids research*. - 2020. - T. 48. - №. W1. - C. W395-W402.
133. Jha U. C. et al. Long non-coding RNAs: emerging players regulating plant abiotic stress response and adaptation // *BMC Plant Biology*. - 2020. - T. 20. - №. 1. - C. 466.
134. Jiang M. et al. PlantPan: A comprehensive multi-species plant pan-genome database // *The Plant Journal*. - 2025. - T. 122. - №. 1. - C. e70144.
135. Jiao Y. et al. Improved Maize Reference Genome with Single-Molecule Technologies // *Nature*. - 2017. - T. 546. - №. 7659. - C. 524-527.
136. Jin J. et al. PLncDB V2.0: A Comprehensive Encyclopedia of Plant Long Noncoding RNAs // *Nucleic Acids Research*. - 2021. - T. 49. - №. D1. - C. D1489-D1495.
137. Jin M. et al. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation // *Scientific reports*. - 2016. - T. 6. - №. 1. - C. 18936.

138. Jin M., Liu H., He C., Fu J., Xiao Y., Wang Y., Xie W., Wang G., Yan J. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation // *Scientific Reports*. - 2016. - T. 6. - №. 1. - C. 18936.
139. Jin Q. et al. Long noncoding RNAs: emerging roles in pulmonary hypertension // *Heart Failure Reviews*. - 2020. - T. 25. - №. 5. - C. 795-815.
140. Joshi A., Romanowska J. Recent advances in computational-based approaches in epigenetics studies // *Epigenetics Methods*. - 2020. - C. 569-590.
141. Joshi R. K. et al. Genome Wide Identification and Functional Prediction of Long Non-Coding RNAs Responsive to *Sclerotinia Sclerotiorum* Infection in *Brassica Napus* // *PLOS ONE*. - 2016. - T. 11. - №. 7. - C. e0158784.
142. Julien P. et al. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation // *PLoS biology*. - 2012. - T. 10. - №. 5. - C. e1001328.
143. Jung Y., Han D. BWA-MEME: BWA-MEM emulated with a machine learning approach // *Bioinformatics*. - 2022. - T. 38. - №. 9. - C. 2404-2413.
144. Köster J., Rahmann S. Snakemake—a scalable bioinformatics workflow engine // *Bioinformatics*. - 2012. - T. 28. - №. 19. - C. 2520-2522.
145. Kallen A. N. et al. The imprinted H19 lncRNA antagonizes let-7 microRNAs // *Molecular cell*. - 2013. - T. 52. - №. 1. - C. 101-112.
146. Kang Y.-J. et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features // *Nucleic acids research*. - 2017. - T. 45. - №. W1. - C. W12-W16.
147. Kapusta A. et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs // *PLoS genetics*. - 2013. - T. 9. - №. 4. - C. e1003470.
148. Karlik E., Ari S., Gozukirmizi N. LncRNAs: genetic and epigenetic effects in plants // *Biotechnology & Biotechnological Equipment*. - 2019. - T. 33. - №. 1. - C. 429-439.
149. Kaur H., Shannon L. M., Samac D. A. A stepwise guide for pangenome development in crop plants: an alfalfa (*Medicago sativa*) case study // *BMC Genomics*. - 2024. - T. 25. - №. 1. - C. 1022.

150. Kazemzadeh M., Safaralizadeh R., Orang A. V. LncRNAs: emerging players in gene regulation and disease pathogenesis // *Journal of Genetics*. - 2015. - T. 94. - №. 4. - C. 771-784.
151. Kertesz M. et al. Genome-wide measurement of RNA secondary structure in yeast // *Nature*. - 2010. - T. 467. - №. 7311. - C. 103-107.
152. Khorkova O., Hsiao J., Wahlestedt C. Basic biology and therapeutic implications of lncRNA // *Advanced drug delivery reviews*. - 2015. - T. 87. - C. 15-24.
153. Kim D. N. et al. Zinc-finger protein CNBP alters the 3-D structure of lncRNA Braveheart in solution // *Nature communications*. - 2020. - T. 11. - №. 1. - C. 148.
154. Kim D. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype // *Nature biotechnology*. - 2019. - T. 37. - №. 8. - C. 907-915.
155. Kim D., Langmead B., Salzberg S. L. HISAT: A Fast Spliced Aligner with Low Memory Requirements // *Nature Methods*. - 2015. - T. 12. - №. 4. - C. 357-360.
156. Kim E.-D., Sung S. Long noncoding RNA: unveiling hidden layer of gene regulatory networks // *Trends in plant science*. - 2012. - T. 17. - №. 1. - C. 16-21.
157. Kim J. Y. et al. Dynamic landscape of long noncoding RNAs during leaf aging in *Arabidopsis* // *Frontiers in Plant Science*. - 2022. - T. 13. - C. 1068163.
158. Kim J. et al. Functional characterization of phytochrome interacting factor 3 in phytochrome-mediated light signal transduction // *The Plant Cell*. - 2003. - T. 15. - №. 10. - C. 2399-2407.
159. Kim T.-K. et al. Widespread transcription at neuronal activity-regulated enhancers // *Nature*. - 2010. - T. 465. - №. 7295. - C. 182-187.
160. Kim T.-K., Hemberg M., Gray J. M. Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers // *Cold Spring Harbor perspectives in biology*. - 2015. - T. 7. - №. 1. - C. a018622.

161. Kimura T. et al. Stabilization of human interferon- α 1 mRNA by its antisense RNA // Cellular and Molecular Life Sciences. - 2013. - T. 70. - №. 8. - C. 1451-1467.
162. Kindgren P., Ivanov M., Marquardt S. Native Elongation Transcript Sequencing Reveals Temperature Dependent Dynamics of Nascent RNAPII Transcription in Arabidopsis // Nucleic Acids Research. - 2020. - T. 48. - №. 5. - C. 2332-2347.
163. Klapproth C. et al. Common Features in lncRNA Annotation and Classification: A Survey // Non-Coding RNA. - 2021. - T. 7. - №. 4. - C. 77.
164. Kong L. et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine // Nucleic acids research. - 2007. - T. 35. - №. 2. - C. W345-W349.
165. Kopp F., Mendell J. T. Functional classification and experimental dissection of long noncoding RNAs // Cell. - 2018. - T. 172. - №. 3. - C. 393-407.
166. Kornblihtt A. R. Chromatin, transcript elongation and alternative splicing // Nature structural & molecular biology. - 2006. - T. 13. - №. 1. - C. 5-7.
167. Kornienko A. E. et al. Population-level annotation of lncRNAs in Arabidopsis reveals extensive expression variation associated with transposable element-like silencing // The Plant Cell. - 2024. - T. 36. - №. 1. - C. 85-111.
168. Kotake Y. et al. Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15INK4B tumor suppressor gene // Oncogene. - 2011. - T. 30. - №. 16. - C. 1956-1962.
169. Kryuchkova-Mostacci N., Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics // Briefings in bioinformatics. - 2018. - T. 18. - №. 2. - C. 205-214.
170. Kung J. T. Y., Colognori D., Lee J. T. Long Noncoding RNAs: Past, Present, and Future // Genetics. - 2013. - T. 193. - №. 3. - C. 651-669.
171. Lai F. et al. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription // Nature. - 2013. - T. 494. - №. 7438. - C. 497-501.

172. Lam M. T. et al. Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription // *Nature*. - 2013. - T. 498. - №. 7455. - C. 511-515.
173. Lamin-Samu A. T. et al. Long non-coding RNA transcriptome landscape of anthers at different developmental stages in response to drought stress in tomato // *Genomics*. - 2022. - T. 114. - №. 4. - C. 110383.
174. Lapierre P., Gogarten J. P. Estimating the size of the bacterial pan-genome // *Trends in genetics*. - 2009. - T. 25. - №. 3. - C. 107-110.
175. Lemmon Z. H. et al. The Role of Cis Regulatory Evolution in Maize Domestication // *PLoS Genetics*. - 2014. - T. 10. - №. 11. - C. e1004745.
176. Li A., Zhang J., Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme // *BMC Bioinformatics*. - 2014. - T. 15. - №. 1. - C. 311.
177. Li J. et al. Cotton pan-genome retrieves the lost sequences and genes during domestication and selection // *Genome Biology*. - 2021. - T. 22. - №. 1. - C. 119.
178. Li L. et al. Genome-wide discovery and characterization of maize long non-coding RNAs // *Genome Biology*. - 2014. - T. 15. - №. 2. - C. R40.
179. Li L., Stoeckert C. J., Roos D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes // *Genome Research*. - 2003. - T. 13. - №. 9. - C. 2178-2189.
180. Li X. et al. lncRNAs: insights into their function and mechanics in underlying disorders // *Mutation Research/Reviews in Mutation Research*. - 2014. - T. 762. - C. 1-21.
181. Li Y. et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits // *Nature biotechnology*. - 2014. - T. 32. - №. 10. - C. 1045-1052.
182. Liu X. et al. Long non-coding RNAs and their biological roles in plants // *Genomics, proteomics & bioinformatics*. - 2015. - T. 13. - №. 3. - C. 137-147.

183. Loewer S. et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells // *Nature genetics*. - 2010. - T. 42. - №. 12. - C. 1113-1117.
184. Lopes R., Agami R., Korkmaz G. GRO-seq, a tool for identification of transcripts regulating gene expression // *Promoter Associated RNA: Methods and Protocols* / под ред. S. Napoli. - New York: Springer New York, 2017. - C. 45-55.
185. Lorenz R. et al. ViennaRNA Package 2.0 // *Algorithms for Molecular Biology*. - 2011. - T. 6. - №. 1. - C. 26.
186. Luco R. F. et al. Regulation of Alternative Splicing by Histone Modifications // *Science*. - 2010. - T. 327. - №. 5968. - C. 996-1000.
187. Ma L., Bajic V. B., Zhang Z. On the classification of long non-coding RNAs // *RNA Biology*. - 2013. - T. 10. - №. 6. - C. 924-933.
188. Ma Y. et al. A pan-transcriptome analysis shows that disease resistance genes have undergone more selection pressure during barley domestication // *BMC Genomics*. - 2019. - T. 20. - №. 1. - C. 12.
189. Malakar P. et al. Long noncoding RNA MALAT1 promotes hepatocellular carcinoma development by SRSF1 upregulation and mTOR activation // *Cancer research*. - 2017. - T. 77. - №. 5. - C. 1155-1167.
190. Mangelsdorf P. C. Corn: its origin, evolution and improvement. - Harvard University Press, 1974.
191. Mangelsdorf P. C., Reeves R. G. The origin of maize // *Proceedings of the National Academy of Sciences*. - 1938. - T. 24. - №. 8. - C. 303-312.
192. Manzano-Morales S. et al. Comparison of gene clustering criteria reveals intrinsic uncertainty in pangenome analyses // *Genome Biology*. - 2023. - T. 24. - №. 1. - C. 250.
193. Marquardt S. et al. Functional consequences of splicing of the antisense transcript COOLAIR on FLC transcription // *Molecular cell*. - 2014. - T. 54. - №. 1. - C. 156-165.

194. Marques A. C., Ponting C. P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness // *Genome Biology*. - 2009. - T. 10. - №. 11. - C. R124.
195. Masuka B., Atlin G. N., Olsen M., Magorokosho C., Labuschagne M., Crossa J., Bänziger M., Pixley K., Vivek B., van Biljon A., Macrobert J., Alvarado G., Prasanna B. M., Makumbi D., Tarekegne A., Das B., Zaman-Allah M., Cairns J. E. Gains in maize genetic improvement in Eastern and Southern Africa. I. CIMMYT hybrid breeding pipeline // *Crop Science*. - 2017. - T. 57. - №. 1. - C. 168-179.
196. Matthews B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme // *Biochimica et Biophysica Acta (BBA)-Protein Structure*. - 1975. - T. 405. - №. 2. - C. 442-451.
197. Matsuoka Y. et al. A single domestication for maize shown by multilocus microsatellite genotyping // *Proceedings of the National Academy of Sciences*. - 2002. - T. 99. - №. 9. - C. 6080-6084.
198. Mattick J. S., Amaral P. P., Carninci P., Carpenter S., Chang H. Y., Chen L. L., Chen R., Dean C., Dinger M. E., Fitzgerald K. A., Gingeras T. R., Guttman M., Hirose T., Huarte M., Johnson R., Kanduri C., Kapranov P., Lawrence J. B., Lee J. T., Mendell J. T., Mercer T. R., Moore K. J., Nakagawa S., Rinn J.L., Spector D. L., Ulitsky I., Wan Y., Wilusz J. E., Wu M. Long non-coding RNAs: definitions, functions, challenges and recommendations // *Nature Reviews Molecular Cell Biology*. - 2023. - T. 24. - №. 6. - C. 430-447.
199. Mattick J. S., Rinn J. L. Discovery and Annotation of Long Noncoding RNAs // *Nature Structural & Molecular Biology*. - 2015. - T. 22. - №. 1. - C. 5-7.
200. Medina C. A., Samac D. A., Yu L.-X. Pan-transcriptome identifying master genes and regulation network in response to drought and salt stresses in Alfalfa (*Medicago sativa* L.) // *Scientific Reports*. - 2021. - T. 11. - №. 1. - C. 17203.
201. Medvedeva Y. A. et al. EpiFactors: a comprehensive database of human epigenetic factors and complexes // *Database*. - 2015. - T. 2015. - C. bav067.

202. Megquier K. et al. BarkBase: Epigenomic Annotation of Canine Genomes // *Genes*. - 2019. - T. 10. - №. 6. - C. 433.
203. Meile L. et al. A fungal avirulence factor encoded in a highly plastic genomic region triggers partial resistance to septoria tritici blotch // *New Phytologist*. - 2018. - T. 219. - №. 3. - C. 1048-1061.
204. Melo C. A. et al. eRNAs are required for p53-dependent enhancer activity and gene transcription // *Molecular cell*. - 2013. - T. 49. - №. 3. - C. 524-535.
205. Merino E. J. et al. RNA Structure Analysis at Single Nucleotide Resolution by Selective 2'-Hydroxyl Acylation and Primer Extension (SHAPE) // *Journal of the American Chemical Society*. - 2005. - T. 127. - №. 12. - C. 4223-4231.
206. Merot-L'anthoene V. et al. Development and evaluation of a genome-wide Coffee 8.5K SNP array and its application for high-density genetic mapping and for investigating the origin of *Coffea arabica* L. // *Plant Biotechnology Journal*. - 2019. - T. 17. - №. 7. - C. 1418-1430.
207. Mirebrahim H., Close T. J., Lonardi S. De novo meta-assembly of ultra-deep sequencing data // *Bioinformatics*. - 2015. - T. 31. - №. 12. - C. i9-i16.
208. Mukherjee N. et al. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability // *Molecular cell*. - 2011. - T. 43. - №. 3. - C. 327-339.
209. Mukhin A. M. et al. RDBMS and NoSQL Based Hybrid Technology for Transcriptome Data Structuring and Processing // *Mathematical Biology and Bioinformatics*. - 2020. - T. 15. - №. 2. - C. 455-470.
210. Muthaiyan M. et al. The Systematic Analysis of LncRNA Expression in Ebola-Infected Human Macrophages. - 2021.
211. Nazipova N. N. Variety of non-coding RNAs in eukaryotic genomes // *Matematicheskaya Biologiya i Bioinformatika*. - 2021. - T. 16. - №. 2. - C. 256-298.
212. Necsulea A. et al. The evolution of lncRNA repertoires and expression patterns in tetrapods // *Nature*. - 2014. - T. 505. - №. 7485. - C. 635-640.

213. Nitsche A., Stadler P. F. Evolutionary clues in lncRNAs // WIREs RNA. - 2017. - T. 8. - №. 1. - C. e1376.
214. Ou S. et al. Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline // Genome Biology. - 2019. - T. 20. - №. 1. - C. 275.
215. Pal D., Rao M. R. S. Long Noncoding RNAs in Pluripotency of Stem Cells and Cell Fate Specification // Long Non Coding RNA Biology / Под ред. M. R. S. Rao. - Singapore: Springer Singapore, 2017. - C. 223-252.
216. Palos K. et al. Identification and functional annotation of long intergenic non-coding RNAs in Brassicaceae // The Plant Cell. - 2022. - T. 34. - №. 9. - C. 3233-3260.
217. Pang L. et al. PMMA quantum dots composites fabricated via use of pre-polymerization // Optics Express. - 2005. - T. 13. - №. 1. - C. 44-49.
218. Paytuví Gallart A. et al. GREENC: A Wiki-Based Database of Plant LncRNAs // Nucleic Acids Research. - 2016. - T. 44. - №. D1. - C. D1161-D1166.
219. Pertea G., Pertea M. GFF Utilities: GffRead and GffCompare // F1000Research. - 2020. - T. 9. - C. 304.
220. Pertea M. et al. StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads // Nature Biotechnology. - 2015. - T. 33. - №. 3. - C. 290-295.
221. Petek M. et al. Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato // Scientific Data. - 2020. - T. 7. - №. 1. - C. 249.
222. Pisignano G., Lodomery M. Epigenetic regulation of alternative splicing: how LncRNAs tailor the message // Non-coding RNA. - 2021. - T. 7. - №. 1. - C. 21.
223. Pisignano G., Pavlaki I., Murrell A. Being in a loop: how long non-coding RNAs organise genome architecture // Essays in Biochemistry. - 2019. - T. 63. - №. 1. - C. 177-186.

224. Ponting C. P., Oliver P. L., Reik W. Evolution and functions of long noncoding RNAs // *Cell*. - 2009. - T. 136. - №. 4. - C. 629-641.
225. Porto F. W., Daulatabad S. V., Janga S. C. Long non-coding RNA expression levels modulate cell-type-specific splicing patterns by altering their interaction landscape with RNA-binding proteins // *Genes*. - 2019. - T. 10. - №. 8. - C. 593.
226. Pozzoli U. et al. Both selective and neutral processes drive GC content evolution in the human genome // *BMC Evolutionary Biology*. - 2008. - T. 8. - №. 1. - C. 99.
227. Prasad D. V., Madhusudanan S., Jaganathan S. uCLUST - a new algorithm for clustering unstructured data // *ARPN Journal of Engineering and Applied Sciences*. - 2015. - T. 10. - №. 5. - C. 2108-2117.
228. Prjibelski A. et al. Using SPAdes De Novo Assembler // *Current Protocols in Bioinformatics*. - 2020. - T. 70. - №. 1. - C. e102.
229. Pronozin A. Y., Afonnikov D. A. ICannoLncRNA : A snakemake pipeline for a long non-coding-RNA search and annotation in transcriptomic sequences // *Genes*. - 2023. - T. 14. - №. 7. - C. 1331.
230. Pronozin A. Y., Bragina M. K., Salina E. A. Crop pangenomes // *Vavilov Journal of Genetics and Breeding*. - 2021. - T. 25. - №. 1. - C. 57.
231. Ramakrishnaiah Y., Kuhlmann L., Tyagi S. Towards a Comprehensive Pipeline to Identify and Functionally Annotate Long Noncoding RNA (LncRNA) // *Computational Biology and Medicine*. - 2020. - T. 127. - C. 104028.
232. Ransohoff J. D., Wei Y., Khavari P. A. The Functions and Unique Features of Long Intergenic Non-Coding RNA // *Nature Reviews Molecular Cell Biology*. - 2018. - T. 19. - №. 3. - C. 143-157.
233. Ranwez V. et al. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons // *Molecular Biology and Evolution*. - 2018. - T. 35. - №. 10. - C. 2582-2584.

234. Rasko D. A. et al. The Pangenome Structure of *Escherichia coli*: Comparative Genomic Analysis of *E. coli* Commensal and Pathogenic Isolates // *Journal of Bacteriology*. - 2008. - T. 190. - №. 20. - C. 6881-6893.
235. Rognes T. et al. VSEARCH: a versatile open source tool for metagenomics // *PeerJ*. - 2016. - T. 4. - C. e2584.
236. Rosenfeld M. G. et al. Altered expression of the calcitonin gene associated with RNA polymorphism // *Nature*. - 1981. - T. 290. - №. 5801. - C. 63-65.
237. Sanchita, Trivedi P. K., Asif M. H. Updates on Plant Long Non-Coding RNAs (LncRNAs): The Regulatory Components // *Plant Cell, Tissue and Organ Culture*. - 2020. - T. 140. - №. 2. - C. 259-269.
238. Sang S. et al. Data integration and evolutionary analysis of long non-coding RNAs in 25 flowering plants // *BMC Genomics*. - 2021. - T. 22. - №. S3. - C. 739.
239. Sasaki Y. The Truth of the F-Measure. - Manchester: MIB-School of Computer Science, University of Manchester, 2007.
240. Schön P. Atomic force microscopy of RNA: State of the art and recent advancements // *Seminars in Cell & Developmental Biology*. - Elsevier, 2018. - C. 209-219.
241. Schulz M. H., Zerbino D. R., Vingron M., Birney E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels // *Bioinformatics*. - 2012. - T. 28. - №. 8. - C. 1086-1092.
242. Sheng L. et al. New insights into the long non-coding RNA SRA: physiological functions and mechanisms of action // *Frontiers in medicine*. - 2018. - T. 5. - C. 244.
243. Shi H. et al. Genome-wide identification of long non-coding RNA for *Botrytis cinerea* during infection to tomato (*Solanum lycopersicum*) leaves // *BMC Genomics*. - 2025. - T. 26. - №. 1. - C. 7.
244. Shiraki T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage //

- Proceedings of the National Academy of Sciences. - 2003. - T. 100. - №. 26. - C. 15776-15781.
245. Shmakov N. A. Improving the quality of barley transcriptome de novo assembling by using a hybrid approach for lines with varying spike and stem coloration // Vavilov Journal of Genetics and Breeding. - 2021. - T. 25. - №. 1. - C. 30.
 246. Shukla S. et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing // Nature. - 2011. - T. 479. - №. 7371. - C. 74-79.
 247. Simopoulos C. M., Weretilnyk E. A., Golding G. B. Molecular traits of long non-protein coding RNAs from diverse plant species show little evidence of phylogenetic relationships // G3: Genes, Genomes, Genetics. - 2019. - T. 9. - №. 8. - C. 2511-2520.
 248. Singh A. et al. Global Transcriptome Characterization and Assembly of the Thermophilic Ascomycete Chaetomium Thermophilum // Genes. - 2021. - T. 12. - №. 10. - C. 1549.
 249. Song J.-H., Cao J.-S., Wang C.-G. BcMF11, a novel non-coding RNA gene from Brassica campestris, is required for pollen development and male fertility // Plant Cell Reports. - 2013. - T. 32. - №. 1. - C. 21-30.
 250. Sousa C. et al. Translational and Structural Requirements of the Early Nodulin Gene enod40, a Short-Open Reading Frame-Containing RNA, for Elicitation of a Cell-Specific Growth Response in the Alfalfa Root Cortex // Molecular and Cellular Biology. - 2001. - T. 21. - №. 1. - C. 354-366.
 251. Staff S. Using the SRA Toolkit to Convert .sra Files into Other Formats. - National Center for Biotechnology Information (US), 2011.
 252. Subramanian S., Kumar S. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome // Genetics. - 2004. - T. 168. - №. 1. - C. 373-381.
 253. Sun K. et al. ISeeRNA: Identification of Long Intergenic Non-Coding RNA Transcripts from Transcriptome Sequencing Data // BMC Genomics. - 2013. - T. 14. - №. S2. - C. S7.

254. Sun L. et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts // *Nucleic acids research*. - 2013. - T. 41. - №. 17. - C. e166.
255. Svergun D. I., Koch M. H. Small-angle scattering studies of biological macromolecules in solution // *Reports on Progress in Physics*. - 2003. - T. 66. - №. 10. - C. 1735.
256. Szcześniak M. W. et al. CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs // *Plant Long Non-Coding RNAs*. - Springer, 2019. - C. 415-429.
257. Tahir Ul Qamar M. et al. ppsPCP: a plant presence/absence variants scanner and pan-genome construction pipeline // *Bioinformatics*. - 2019. - T. 35. - №. 20. - C. 4156-4158.
258. Talyan S. et al. CALINCA—a novel pipeline for the identification of lncRNAs in podocyte disease // *Cells*. - 2021. - T. 10. - №. 3. - C. 692.
259. Tang H. et al. Resilience of Maize to Environmental Stress: Insights into Drought and Heat Tolerance // *International Journal of Molecular Sciences*. - 2025. - T. 26. - №. 11. - C. 5274.
260. Tettelin H. et al. Comparative genomics: the bacterial pan-genome // *Current opinion in microbiology*. - 2008. - T. 11. - №. 5. - C. 472-477.
261. Thomson D. W., Dinger M. E. Endogenous microRNA sponges: evidence and controversy // *Nature Reviews Genetics*. - 2016. - T. 17. - №. 5. - C. 272-283.
262. Tranchant-Dubreuil C., Rouard M., Sabot F. Plant pangenome: impacts on phenotypes and evolution // *Annual Plant Reviews*. - 2019.
263. Trapnell C. et al. Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks // *Nature Protocols*. - 2012. - T. 7. - №. 3. - C. 562-578.
264. Traubenik S., Charon C., Blein T. From environmental responses to adaptation: the roles of plant lncRNAs // *Plant Physiology*. - 2024. - T. 195. - №. 1. - C. 232-244.

265. Tripathi V. et al. The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation // *Molecular cell*. - 2010. - T. 39. - №. 6. - C. 925-938.
266. Tsai M.-C. et al. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes // *Science*. - 2010. - T. 329. - №. 5992. - C. 689-693.
267. Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs // *Nature Reviews Genetics*. - 2016. - T. 17. - №. 10. - C. 601-614.
268. Ulitsky I., Bartel D. P. lincRNAs: genomics, evolution, and mechanisms // *Cell*. - 2013. - T. 154. - №. 1. - C. 26-46.
269. Unver T., Tombuloglu H. Barley long non-coding RNAs (lncRNA) responsive to excess boron // *Genomics*. - 2020. - T. 112. - №. 2. - C. 1947-1955.
270. Van Bel M. et al. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics // *Nucleic acids research*. - 2018. - T. 46. - №. 1. - C. D1190-D1196.
271. Venturini L. et al. Leveraging multiple transcriptome assembly methods for improved gene structure annotation // *GigaScience*. - 2018. - T. 7. - №. 8. - C. giy093.
272. Veras A. et al. Pan4Draft: a computational tool to improve the accuracy of pan-genomic analysis using draft genomes // *Scientific reports*. - 2018. - T. 8. - №. 1. - C. 9670.
273. Vernikos G. et al. Ten years of pan-genome analyses // *Current opinion in microbiology*. - 2015. - T. 23. - C. 148-154.
274. Walley J. W. et al. Integration of Omic Networks in a Developmental Atlas of Maize // *Science*. - 2016. - T. 353. - №. 6301. - C. 814-818.
275. Wang A. et al. Genome-wide analysis of long non-coding RNAs unveils the regulatory roles in the heat tolerance of Chinese cabbage (*Brassica rapa* ssp. *chinensis*) // *Scientific reports*. - 2019. - T. 9. - №. 1. - C. 5002.
276. Wang H. L. V., Chekanova J. A. An overview of methodologies in studying lncRNAs in the high-throughput era: when acronyms ATTACK! //

- Plant long Non-Coding RNAs: Methods and Protocols / под ред. J. A. Chekanova, H. L. V. Wang. - New York: Springer New York, 2019. - С. 1-30.
277. Wang H. et al. Analysis of non-coding transcriptome in rice and maize uncovers roles of conserved lncRNAs associated with agriculture traits // *The Plant Journal*. - 2015. - Т. 84. - №. 2. - С. 404-416.
 278. Wang J. et al. Genome-wide analysis of tomato long non-coding RNAs and identification as endogenous target mimic for microRNA in response to TYLCV infection // *Scientific reports*. - 2015. - Т. 5. - №. 1. - С. 16946.
 279. Wang J. et al. Non-coding RNAs and their roles in stress response in plants // *Genomics, proteomics & bioinformatics*. - 2017. - Т. 15. - №. 5. - С. 301-312.
 280. Wang J. et al. Transcriptome analysis during vernalization in wheat (*Triticum aestivum* L.) // *BMC Genomic Data*. - 2023. - Т. 24. - №. 1. - С. 43.
 281. Wang K. C., Chang H. Y. Molecular mechanisms of long noncoding RNAs // *Molecular cell*. - 2011. - Т. 43. - №. 6. - С. 904-914.
 282. Wang L. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model // *Nucleic acids research*. - 2013. - Т. 41. - №. 6. - С. e74.
 283. Wang P. et al. Global Profiling of lncRNAs Expression Responsive to Allopolyploidization in *Cucumis* // *Genes*. - 2020. - Т. 11. - №. 12. - С. 1500.
 284. Wang P. et al. Identification and functional prediction of cold-related long non-coding RNA (lncRNA) in grapevine // *Scientific Reports*. - 2019. - Т. 9. - №. 1. - С. 6638.
 285. Wang T. et al. Novel phosphate deficiency-responsive long non-coding RNAs in the legume model plant *Medicago truncatula* // *Journal of experimental botany*. - 2017. - Т. 68. - №. 21-22. - С. 5937-5948.
 286. Wang X. et al. Noncanonical transcription initiation is primarily tissue specific and epigenetically tuned in paleopolyploid plants // *Plant Cell*. - 2024. - Т. 37. - С. koae288.

287. Wang X. et al. Research progress on the roles of lncRNAs in plant development and stress responses // *Frontiers in Plant Science*. - 2023. - T. 14. - C. 1138901.
288. Wang Y. et al. Identification, stability and expression of Sirt1 antisense long non-coding RNA // *Gene*. - 2014. - T. 539. - №. 1. - C. 117-124.
289. Wang Y. et al. LncDLSM: identification of long non-coding RNAs with deep learning-based sequence model // *IEEE Journal of Biomedical and Health Informatics*. - 2023. - T. 27. - №. 4. - C. 2117-2127.
290. Wences A. H., Schatz M. C. Metassembler: merging and optimizing de novo genome assemblies // *Genome Biology*. - 2015. - T. 16. - №. 1. - C. 207.
291. Wolfe K. H., Sharp P. M. Mammalian gene evolution: Nucleotide sequence divergence between mouse and rat // *Journal of Molecular Evolution*. - 1993. - T. 37. - №. 4. - C. 441-456.
292. Wright E. Accurately clustering biological sequences in linear time by relatedness sorting // *Nature Communications*. - 2024. - T. 15. - №. 1. - C. 3047.
293. Wu H.-J. et al. Widespread long noncoding RNAs as endogenous target mimics for microRNAs in plants // *Plant physiology*. - 2013. - T. 161. - №. 4. - C. 1875-1884.
294. Wu J. et al. A Novel Hypoxic Stress-Responsive Long Non-Coding RNA Transcribed by RNA Polymerase III in Arabidopsis // *RNA Biology*. - 2012. - T. 9. - №. 3. - C. 302-313.
295. Wu T. D., Watanabe C. K. GMAP: A Genomic Mapping and Alignment Program for mRNA and EST Sequences // *Bioinformatics*. - 2005. - T. 21. - №. 9. - C. 1859-1875.
296. Xiao S.-J. et al. TiSGeD: a database for tissue-specific genes // *Bioinformatics*. - 2010. - T. 26. - №. 9. - C. 1273-1275.
297. Xie C. et al. NONCODEv4: exploring the world of long non-coding RNA genes // *Nucleic acids research*. - 2014. - T. 42. - №. D1. - C. D98-D103.
298. Xie Y. et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads // *Bioinformatics*. - 2014. - T. 30. - №. 12. - C. 1660-1666.

299. Yanai I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification // *Bioinformatics*. - 2005. - T. 21. - №. 5. - C. 650-659.
300. Yang C. et al. LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning // *Bioinformatics*. - 2018. - T. 34. - №. 22. - C. 3825-3834.
301. Yang G. et al. Population transcriptomic analysis identifies the comprehensive lncRNAs landscape of spike in wheat (*Triticum aestivum* L.) // *BMC Plant Biology*. - 2022. - T. 22. - №. 1. - C. 450.
302. Yang L., Froberg J. E., Lee J. T. Long noncoding RNAs: fresh perspectives into the RNA world // *Trends in biochemical sciences*. - 2014. - T. 39. - №. 1. - C. 35-43.
303. Yang Z. Among-site rate variation and its impact on phylogenetic analyses // *Trends in ecology & evolution*. - 1996. - T. 11. - №. 9. - C. 367-372.
304. Yang Z., Cao Y., Shi Y., Qin F., Jiang C., Yang S. Genetic and molecular exploration of maize environmental stress resilience: Toward sustainable agriculture // *Molecular Plant*. - 2023. - T. 16. - №. 10. - C. 1496-1517.
305. Yang Z. et al. LncRNA expression profile and ceRNA analysis in tomato during flowering // *PLoS One*. - 2019. - T. 14. - №. 1. - C. e0210650.
306. Yao W. et al. Exploring the rice dispensable genome using a metagenome-like assembly strategy // *Genome Biology*. - 2015. - T. 16. - №. 1. - C. 187.
307. Ye X. et al. Role of lncRNAs in cis- and trans- regulatory responses to salt in *Populus trichocarpa* // *The Plant Journal*. - 2022. - T. 110. - №. 4. - C. 978-993.
308. Yi X., Zhang Z., Ling Y., Xu W., Su Z. PNRD: A Plant Non-Coding RNA Database // *Nucleic Acids Research*. - 2015. - T. 43. - №. D1. - C. D982-D989.
309. Yin J. et al. Identification and functional deciphering suggested the regulatory roles of long intergenic ncRNAs (lincRNAs) in increasing grafting pepper resistance to *Phytophthora capsici* // *BMC Genomics*. - 2021. - T. 22. - №. 1. - C. 868.

310. Yoon J.-H. et al. Scaffold function of long non-coding RNA HOTAIR in protein ubiquitination // *Nature communications*. - 2013. - T. 4. - №. 1. - C. 2939.
311. Yu X. et al. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues // *Nucleic acids research*. - 2006. - T. 34. - №. 17. - C. 4925-4936.
312. Yu Y. et al. The role of wheat long intergenic non-coding RNAs in drought stress // *Crop and Pasture Science*. - 2025. - T. 76. - №. 6.
313. Yuan D., Ahamed A., Burgin J., Cummins C., Devraj R., Gueye K., Cochrane G. The European nucleotide archive in 2023 // *Nucleic Acids Research*. - 2024. - T. 52. - №. D1. - C. D92-D97.
314. Zerbino D. R., Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs // *Genome research*. - 2008. - T. 18. - №. 5. - C. 821-829.
315. Zhang P. et al. Non-Coding RNAs and their Integrated Networks // *Journal of Integrative Bioinformatics*. - 2019. - T. 16. - C. 20190027.
316. Zhang P. et al. Plant miRNA-lncRNA Interaction Prediction with the Ensemble of CNN and IndRNN // *Interdisciplinary Sciences: Computational Life Sciences*. - 2019. - T. 12. - №. 1. - C. 82-89.
317. Zhang X. et al. Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels // *International Journal of Molecular Sciences*. - 2019. - T. 20. - №. 22. - C. 5573.
318. Zhang X. et al. The long non-coding RNA lncRNA973 is involved in cotton response to salt stress // *BMC Plant Biology*. - 2019. - T. 19. - №. 1. - C. 459.
319. Zhang Y.-C. et al. Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice // *Genome Biology*. - 2014. - T. 15. - №. 12. - C. 512.
320. Zhang Z. KaKs_Calculator 3.0: calculating selective pressure on coding and non-coding sequences // *Genomics, Proteomics and Bioinformatics*. - 2022. - T. 20. - №. 3. - C. 536-540.

321. Zhao M. et al. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives // BMC Bioinformatics. - 2013. - T. 14. - №. S11. - C. S1.
322. Zhao Q. et al. LncPipe: A Nextflow-based pipeline for identification and analysis of long non-coding RNAs from RNA-Seq data // J Genet Genomics. - 2018. - T. 45. - №. 7. - C. 399-401.
323. Zhao X. et al. Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA // Nature communications. - 2018. - T. 9. - №. 1. - C. 5056.
324. Zhao Y. et al. PGAP: pan-genomes analysis pipeline // Bioinformatics. - 2012. - T. 28. - №. 3. - C. 416-418.
325. Zhao Y. et al. PanGP: a tool for quickly analyzing bacterial pan-genome profile // Bioinformatics. - 2014. - T. 30. - №. 9. - C. 1297-1299.
326. Zhao Z. et al. Long non-coding RNAs: new players in plants // International Journal of Molecular Sciences. - 2022. - T. 23. - №. 16. - C. 9301.
327. Zhou B. et al. EVLncRNAs 2.0: An Updated Database of Manually Curated Functional Long Non-Coding RNAs Validated by Low-Throughput Experiments // Nucleic Acids Research. - 2021. - T. 49. - №. D1. - C. D86-D91.
328. Zhou B. et al. EVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput experiments // Nucleic Acids Research. - 2018. - T. 46. - №. D1. - C. D100-D105.
329. Zhong Y. et al. Pan-transcriptomic analysis reveals alternative splicing control of cold tolerance in rice // The Plant Cell. - 2024. - T. 36. - №. 6. - C. 2117-2139.
330. Zhu J., Liu M., Liu X., Dong Z. RNA polymerase II activity revealed by GRO-seq and pNET-seq in Arabidopsis // Nature Plants. - 2018. - T. 4. - №. 12. - C. 1112-1123.
331. Zhu Q.-H., Wang M.-B. Molecular functions of long non-coding RNAs in plants // Genes. - 2012. - T. 3. - №. 1. - C. 176-190.

332. Zhu Y. et al. Revealing the novel complexity of plant long non-coding RNA by strand-specific and whole transcriptome sequencing for evolutionarily representative plant species // BMC Genomics. - 2022. - Т. 23. - №. S4. - С. 381.
333. Ziegler C., Kretz M. The more the merrier—complexity in long non-coding RNA loci // Frontiers in Endocrinology. - 2017. - Т. 8. - С. 90.
334. Zimin A. V. et al. The MaSuRCA genome assembler // Bioinformatics. - 2013. - Т. 29. - №. 21. - С. 2669-2677. Żmieńko A. et al. Copy number polymorphism in plant genomes // Theoretical and Applied Genetics. - 2014. - Т. 127. - №. 1. - С. 1-18.
335. Zorin E. A. et al. Analysis of alternative splicing events in the root tips and nodules of *Pisum sativum* L. // Ecological genetics. - 2019. - Т. 17. - №. 1. - С. 53-63.
336. Бейлерли О. А., Гареев И. Ф. Длинные некодирующие РНК: какие перспективы? // Профилактическая медицина. - 2020. - Т. 23. - №. 2. - С. 124-128.
337. Гордиюк В. В. Длинные некодирующие РНК - камертон в регуляции клеточных процессов // Ukrainian Biochemical Journal. - 2014. - Т. 86. - №. 2. - С. 5-15.
338. Канукова К. Р. и др. ДНК-маркеры в растениеводстве // Известия Кабардино-Балкарского научного центра РАН. - 2019. - №. 6 (92). - С. 220-232.
339. Коробка А. Н., Орленко С. Ю., Алексеенко Е. В., Малышева Н. Н. и др. Система земледелия Краснодарского края на агроландшафтной основе. - Краснодар, 2015. - 352 с.
340. Лукина С. С. и др. Роль длинных некодирующих РНК и ДНК-метилирования в патогенезе рака яичников // Патологическая физиология и экспериментальная терапия. - 2022. - Т. 66. - №. 4. - С. 143-156.
341. Назипова Н. Н. Разнообразие некодирующих РНК в геномах эукариот // Математическая биология и биоинформатика. - 2021. - Т. 16. - №. 2. - С. 256-298.

342. Полетаев Г. М., Воловик В. Ф., Гончаров В. М. и др. Новые сорта и гибриды кукурузы и сорговых культур // Кукуруза и сорго. - 2008. - №. 3. - С. 8.
343. Сотченко В. С., Багринцева В. Н., Сотченко Е. Ф., Горбачева А. Г., Ревякин Е. Л. Перспективная ресурсосберегающая технология производства кукурузы на зерно: Методические рекомендации. - ФГНУ «Росинформагротех», 2009. - 72 с.
344. Сохранить и приумножить на практике кукуруза, рис, пшеница. Практическое руководство по устойчивому производству зерновых. - ООН продовольственная и сельскохозяйственная организация. - Рим, 2016. - 124 с.
345. Стулин А. Ф. Продуктивность гибридов кукурузы в зависимости от густоты растений и уровня минерального питания // Кукуруза и сорго. - 2009. - №. 1. - С. 4.
346. Татосян К. А. и др. Функциональные особенности длинных некодирующих РНК, содержащих последовательности мобильных генетических элементов // Молекулярная биология. - 2020. - Т. 54. - №. 5. - С. 718-724.
347. Шиндин А. П., Бгринцева В. Н., Борщ Т. И., Горбачев А. Г., Сотченко В. С., Сотченко Е. Ф., Сотченко Ю. В. Кукуруза. Современная технология выращивания. - М., 2009. - 127 с.
348. Шпаар Д., Гинапп К. и др. Кукуруза: выращивание, уборка, консервирование и использование: Учебно-практическое руководство. - М.: ДЛВ Агродело, 2006. - 390 с.