На правах рукописи

Шмаков Николай Александрович

ПОИСК ГЕНОВ, АССОЦИИРОВАННЫХ С ЧАСТИЧНЫМ АЛЬБИНИЗМОМ И МЕЛАНИЗМОМ У ЯЧМЕНЯ *HORDEUM VULGARE* L., НА ОСНОВЕ АНАЛИЗА ТРАНСКРИПТОМНЫХ ДАННЫХ

1.5.8. Математическая биология, биоинформатика
1.5.7. Генетика

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата биологических наук

Новосибирск – 2023

Работа выполнена в лаборатории эволюционной биоинформатики и теоретической генетики Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», г. Новосибирск.

Научные руководители:	**Афонников Дмитрий Аркадьевич** к.б.н., доцент, зав. лабораторией эволюционной биоинформатики и теоретической генетики ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики СО РАН», г. Новосибирск **Хлесткина Елена Константиновна** д.б.н., профессор РАН, директор Федерального исследовательского центра «Всероссийский институт генетических ресурсов растений имени Н.И. Вавилова», г. Санкт-Петербург
Официальные оппоненты:	**Голденкова-Павлова Ирина Васильевна -** д.б.н., доцент, в.н.с., руководитель группы функциональной геномики ФГБУН Институт физиологии растений им. К.А. Тимирязева РАН, г. Москва **Касьянов Артем Сергеевич -** к.ф.-м.н., доцент, с.н.с. лаборатории геномики растений ФГБУ науки Институт проблем передачи информации им. А.А. Харкевича РАН, г. Москва
Ведущее учреждение:	ФГБУН Институт общей генетики им. Н.И.Вавилова РАН, г. Москва

Защита диссертации состоится « »_____ 2024г. на утреннем заседании диссертационного совета 24.1.239.01 на базе ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики СО РАН», в конференц-зале Института по адресу:
пр. академика Лаврентьева 10, г. Новосибирск, 630090,
тел.: (383) 363-49-06 (1321); e-mail: dissov@bionet.nsc.ru

С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте института www.icgbio.ru
Автореферат разослан « »_____ 202__ г.

Учёный секретарь
диссертационного совета,
доктор биологических наук Хлебодарова Т. М.

Общая характеристика работы

Актуальность темы исследования. Растительные пигменты – собирательное название для веществ, придающих окраску различным органам растений. Пигменты подразделяются на разные типы в соответствии со своей химической структурой. Их биосинтез осуществляется за счет функционирования генов, формирующих метаболические сети. Разные пигменты обеспечивают различные типы окраски.

Наиболее активно изучаются среди растительных пигментов хлорофиллы – производные порфирина, участвующие в процессе фотосинтеза [Wu, 2018]. Недостаток хлорофилла является следствием метаболических нарушений в результате дисфункции генов его биосинтеза и проявляются как альбинизм. Растения с частичным альбинизмом являются перспективным объектом для изучения особенностей биосинтеза и распределения хлорофилла в растительных органах и тканях. У ячменя *Hordeum vulgare* L. один из фенотипов демонстрирует такое нарушение пигментации – альбинизм колосковой чешуи, леммы (цветковой чешуи), ушек и частично стебля. Этот фенотип обусловлен мутацией в гене *Alm*, локализованном в коротком плече хромосомы 3H ячменя [Nonaka, 1973]. Структура и функции гена *Alm* на данный момент неизвестны. Мутантная форма гена *Alm* представляет собой подходящую модель для изучения механизмов координированной работы ядерных и пластидных генов.

Другим важным типом растительных пигментов являются меланины и подобные им соединения. Именно меланины обеспечивают черную окраску органов растений. У ячменя формирование черной окраски колосковой и цветковой чешуй и перикарпа ячменя обусловлено доминантным аллелем гена *Blp*, который локализован в длинном плече хромосомы 1H [Buckley, 1930]. Его функция также на данный момент не известна. Данный ген может стать подходящей моделью для выявления методами функциональной геномики и транскриптомики ранее неизученных метаболических и генных сетей растений, к каким относится путь синтеза меланин-подобных пигментов.

Цель и задачи

Цель работы – выявление генов, ассоциированных с формированием частичного альбинизма и меланизма у колоса ячменя, на основе биоинформатического анализа транскриптомов сорта Bowman и почти изогенных линий i:Bw*Alm* и i:Bw*Blp*, выявление функциональных особенностей этих генов и их роли в метаболических путях биосинтеза пигментов.

Для достижения данной цели были поставлены следующие задачи:

1. Сформировать вычислительные конвейеры с использованием программ биоинформатического анализа транскриптомных данных RNA-seq для реконструкции и анализа последовательностей транскриптов.
2. Разработать подход для оценки качества конвейеров программ, используемых для анализа данных RNA-seq, и найти оптимальные параметры для обработки транскриптомных библиотек ячменя.

3. Определить гены, экспрессирующиеся дифференциально у ячменя сорта Bowman и линий i:Bw*Alm* и i:Bw*Blp*, провести функциональный анализ полученных дифференциально экспрессирующихся генов, выявить термины генной онтологии и метаболические пути, статистически значимо обогащённые для этих генов.
4. Провести поиск и функциональный анализ транскриптов, обнаруженных в транскриптоме исследуемых линий, но не аннотированных ранее в геноме ячменя.

Научная новизна. В работе впервые был проведён транскриптомный анализ почти изогенных линий ячменя, контрастных по окраске колоса. Гены, понижающие экспрессию в лемме ячменя линии i:Bw*Alm*, характеризующейся альбинизмом колоса, по сравнению с леммой ячменя сорта Bowman, взятого в качестве контроля, связаны с аэробным дыханием и фотодыханием. Гены, повышающие экспрессию в этой линии, связаны с протеолизом и защитным ответом. В транскриптоме линии i:Bw*Alm* был обнаружен транскрипт, отсутствующий в транскриптоме леммы ячменя сорта Bowman, кодирующий белковый продукт, содержащий домены прохибитина. Независимая экспериментальная проверка показала, что ген, кодирующий этот транскрипт, локализован в коротком плече хромосомы 3H ячменя.

Гены, повышающие экспрессию в линии ячменя i:Bw*Blp*, характеризующейся меланизмом колоса, по сравнению с сортом Bowman, задействованы в метаболизме жирных кислот, фенольных соединений и полихинонов. Гены, понижающие экспрессию в этой линии, участвуют в биосинтезе хлорофилла и фотосинтезе. Гены, локализованные в пластидном геноме, понижают свою экспрессию в лемме ячменя линии i:Bw*Blp*, причём гены, кодирующие белки, участвующие в фотосинтезе, понижают экспрессию сильнее, чем гены рибосомных РНК.

Использовании нескольких конвейеров биоинформатической обработки библиотек RNA-seq с последующим выбором наиболее оптимального конвейера для имеющихся данных с коррекцией на экспериментально проверенные уровни изменения экспрессии для ряда тестовых генов позволяет достичь большей точности и чувствительности в определении дифференциальной экспрессии генов. Использование нескольких сборщиков транскриптома *de novo* и последующее объединение полученных сборок в одну общую гибридную сборку повышает точность и чувствительность в определении последовательностей транскриптов.

Теоретическая и практическая значимость работы. В работе показана важность использования множественных конвейеров для биоинформатической обработки RNA-seq с последующим отбором наиболее оптимального конвейера по ряду характеристик. Это позволяет получить более точные оценки дифференциальной экспрессии генов и их изоформ. Также показана важность использования множественных сборщиков транскриптома *de novo* с последующей компоновкой полученных результатов

в одну гибридную сборку транскриптома *de novo*, что повышает точность определения структуры транскриптов.

В работе наблюдается изменение экспрессии генов в лемме ячменя линии i:Bw*Alm*, характеризующейся частичным альбинизмом, по сравнению с сортом Bowman, и показано участие генов, повышающих экспрессию в линии i:Bw*Alm*, в защитном ответе и протеолизе. Для генов, понижающих экспрессию в этой линии, показана связь с синтезом хлорофилла и фотосинтезом. В транскриптоме линии i:Bw*Alm* обнаружен транскрипт, не представленный в транскриптоме сорта Bowman. Ген, кодирующий этот транскрипт, локализован в коротком плече хромосомы ячменя 3H.

Был проведён анализ транскриптома ячменя почти изогенной линии i:Bw*Blp*, характеризующейся меланизмом колоса. Гены, повышающие экспрессию в лемме ячменя линии i:Bw*Blp* по сравнению с сортом Bowman, участвуют в метаболизме фенилпропаноидов и жирных кислот. Гены, понижающие экспрессию в этой линии, участвуют в биосинтезе хлорофилла и фотосинтезе. Данные результаты позволяют предположить участие пластид в процессе синтеза меланинов в клетках леммы ячменя.

Положения, выносимые на защиту

1. Предложен метод оптимизации вычислительного конвейера для биоинформатического анализа экспериментов RNA-seq, повышающий точность оценки дифференциальной активности генов, который основан на использовании данных независимой верификации изменения экспрессии генов с помощью ОТ-ПЦР.
2. Формирование частичного дефицита хлорофилла в колосе ячменя (*Hordeum vulgare* L.) мутантной линии i:Bw*Alm* сопровождается понижением уровня экспрессии генов фотосинтеза, аэробного дыхания и усвоения азота, а также активацией в клетках оболочки зерновки гена, локализованного в коротком плече хромосомы 3H и кодирующего белок с доменом прохибитина.
3. Формирование меланиновой окраски колоса ячменя в линии i:Bw*Blp* связано с повышением экспрессии генов в перикарпе зерновки и цветковой чешуе, участвующих в биосинтезе *о*-дихинонов и фенилпропаноидов.

Апробация результатов. Результаты диссертационной работы были представлены на конференциях: PlantGen2017 (Алмата, 2017), Высокопроизводительное Секвенирование в Геномике (Новосибирск, 2017), BGRS-SB (Новосибирск, 2018), конгресс «Биотехнология: состояние и перспективы развития. Науки о жизни» (Москва, 2019), CBB-2019 (Будапешт, 2019).

По результатам диссертационной работы было опубликовано 4 статьи в журналах, индексируемых в базах данных Российский индекс научного цитирования, Scopus и Web of Science.

Структура диссертации. Диссертация состоит из шести разделов: введения, обзора литературы, материалов и методов, результатов, обсуждения результатов, заключения и списка использованной литературы. Текст диссертации изложен на 148 страницах, содержит 22 рисунка и 18 таблиц.

Личный вклад автора. Автором диссертации был проведён биоинформатический анализ библиотек коротких прочтений: фильтрация, картирование, подсчёт уровней экспрессии, поиск и функциональный анализ дифференциальной экспрессии, реконструкция транскриптома *de novo*, интерпретация полученных результатов.

Содержание работы

Глава 1. Обзор литературы

В обзоре литературы рассмотрена структура генома и транскриптома ячменя, известные молекулярные механизмы формирования окраски колоса ячменя. Рассмотрены современные методы транскриптомных исследований, в частности, массовое высокопроизводительное секвенирование РНК (RNA-seq), биоинформатические методы анализа данных RNA-seq.

Глава 2. Методология работы

Использованные материалы. В работе использованы линия ячменя i:Bw*Alm*, характеризующаяся альбинизмом колоса (идентификатор коллекции Nordic GenBank – NGB 20419), линия i:Bw*Blp*, характеризующаяся меланизмом колоса (NGB 20470), и сорт Bowman (NGB 22812), взятый в качестве контроля к обеим линиям. Растения были выращены в ЦКП Лаборатория искусственного выращивания растений ИЦиГ СО РАН Генераловой Г.В. и Кукоевой Т.В. Выделение РНК из биологических образцов проводилось Шоевой О.Ю. и Глаголевой А.Ю. Библиотеки коротких прочтений были секвенированы на платформе IonTorrent PGM в ЦКП Геномика ИЦиГ СО РАН Васильевым Г.В.

Линия i:Bw*Alm* содержит мутацию в гене *Alm*, локализованном в коротком плече хромосомы 3H, в районе, содержащем 229 генов. Линия i:Bw*Blp* содержит мутацию в гене *Blp*, локализованном в длинном плече хромосомы 1H, в районе, в котором находится 21 ген. Последовательности и молекулярные функции генов *Alm* и *Blp* на момент работы над диссертацией были неизвестны.

Анализ транскриптомов линий i:Bw*Alm* и i:Bw*Blp* проводился по отдельности. В обоих случаях в качестве контроля была использован сорт ячменя Bowman. Для анализа транскриптома линии i:Bw*Alm* РНК была выделена из развивающихся колосков. Для анализа транскриптома линии i:Bw*Blp* были взяты лемма и перикарп на ранней стадии восковой спелости.

Компьютерная обработка данных. Биоинформатический анализ данных RNA-seq направлен на выявление дифференциально экспрессирующихся генов (ДЭГ) и может быть подразделён на два основных типа – анализ с использованием референсных последовательностей (основанный на картировании библиотек) и анализ без использования референсных последовательностей (*de novo* реконструкция транскриптома). Анализ на основе картирования библиотек позволяет использовать существующую

функциональную аннотацию генов организма для лучшего понимания изучаемых процессов. Сборка транскриптома *de novo* позволяет получить последовательности транскриптов, которые ранее для исследуемого вида описаны не были. Таким образом, эти методы взаимно дополняют друг друга. В данной работе для анализа транскриптомов двух изучаемых линий ячменя были использованы оба метода обработки.

Анализ на основе картирования включает несколько стадий, на каждой из которых могут быть использованы различные подходы и программные продукты. Однако, нельзя *a priori* утверждать, какие из используемых подходов окажутся наиболее подходящими для конкретных транскриптомных библиотек. Для выбора оптимального подхода выявления ДЭГ для имеющихся данных были сформированы 36 конвейеров, включающих набор различных программ на разных стадиях обработки данных. Схема построения конвейеров биоинформатической обработки представлена на рис. 1.

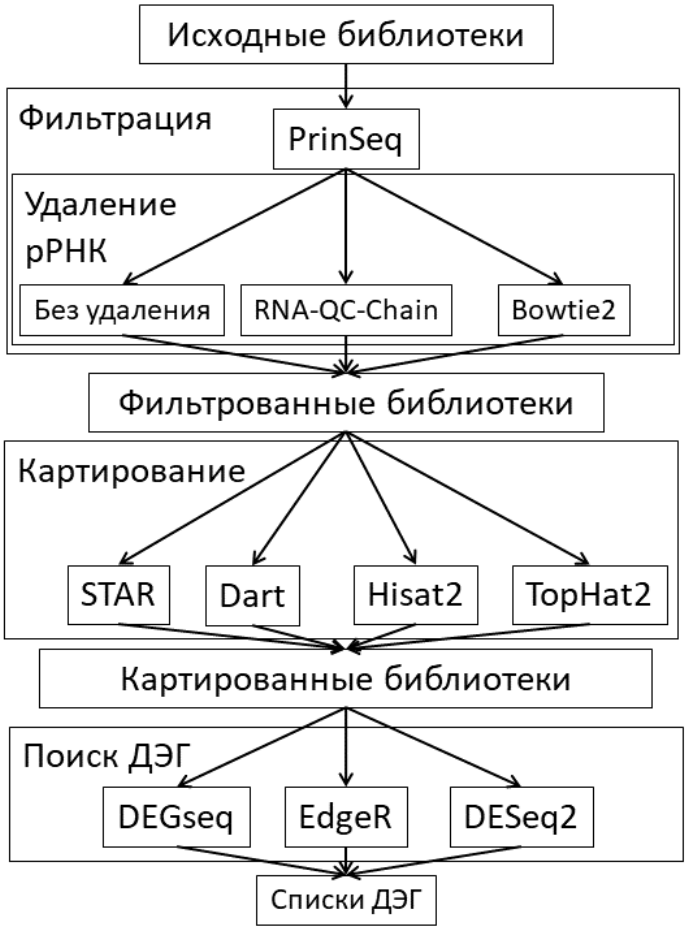

Рисунок 1. Схема выявления списков ДЭГ на основе библиотек RNA-seq с использованием картирования прочтений. Три основных этапа анализа выделены прямоугольниками. Варианты обработки данных на каждом этапе отмечены вложенными прямоугольниками с указанием программы/алгоритма анализа. Последовательность выполнения различных вариантов обработки показана стрелками. Детали использования программ приведены в тексте.

Обработка библиотек. Для оценки качества библиотек использована программа FastQC. Для удаления адаптеров использована программа Cutadapt. Для фильтрации прочтений по длине и качеству использована программа Prinseq-lite.

Удаление рРНК. На этом этапе использовалось три варианта анализа: (1) без фильтрации рРНК, (2) фильтрация с помощью программы RNA-QC-Chain (версия 1.0), (3) фильтрация с помощью картирования картирования

прочтений на последовательности рРНК программой Bowtie2 и удалением выравненных прочтений.

Картирование. Картирование прочтений на референсный геном проводилось четырьмя методами: (1) STAR (вер.2.6.1a), (2) Dart (вер.1.3.2), (3) HISAT2 (вер.2.1.0), (4) TopHat2 (вер. 2.1.1). Существующая сборка генома использованного в работе сорта Bowman не полна и недостаточно аннотирована, поэтому в качестве референса использована сборка генома ячменя сорта Morex версии IBSC v.2 из базы данных Ensembl plants вер. 49.

Идентификация ДЭГ. Для оценки уровня экспрессии генов использовалась программа FeatureCounts из пакета Subread. Гены сровнем экспрессии менее 2 cpm хотя бы в двух образцах из шести, исключались из рассмотрения. ДЭГ определялись на основе трех вариантов анализа: (1) edgeR (вер. 3.20.9, использован точный тест Фишера), (2) DESeq2 (вер.1.18.1, тест отношения правдоподобия), (3) DEGseq (вер.1.32.0, метод MARS). Поправка на множественное сравнение проводилась по методу Бенджамини-Хохберга. Гены, для которых достоверность изменения экспрессии не превышала 0,05 после поправки на множественное сравнение, считаются имеющими достоверную дифференциальную экспрессию

Оценка производительности конвейеров. Все комбинации вариантов обработки данных на каждом этапе анализа дают в общей сложности 3x4x3=36 конвейеров биоинформатической обработки. Полученные с помощью этих конвейеров результаты были оценены по следующим характеристикам: доля картированных прочтений (F_m), доля уникально картированных прочтений (F_u). Для ряда генов была проведена верификация дифференциальной экспрессии с помощью количественной полимеразной реакции в реальном времени (кПЦР). Для всех использованных конвейеров была оценен коэффициент корреляции Пирсона (r) между уровнями изменения экспрессии генов, определёнными биоинформатически, и полученными экспериментально. Робастность полученных коэффициентов корреляции была оценена методом бутстрепа (S_r). Параметры F_m, F_u, r для 36 конвейеров были ранжированы по убыванию, S_r – по возрастанию. Сумма рангов этих параметров оценивала точность работы метода: конвейеры с максимальной суммой считались наиболее оптимальными и были использованы для идентификации и функционального анализа ДЭГ.

Функциональный анализ ДЭГ. Для списка ДЭГ, полученного с помощью конвейера обработки, признанного наиболее эффективным, был проведен функциональный анализ. Для оценки обогащения терминов Генной Онтологии (ГО) использован онлайн-сервис Singular Enrichment Analysis (SEA) базы данных (БД) AgriGO v.2. Для оценки представленности ДЭГ в метаболических путях использована БД PlantCyc v.3 и анализ на основе гипергеометрического распределения. Функциональный анализ генов с пониженной и повышенной экспрессией проводился по отдельности.

Отдельно рассмотрена экспрессия генов, локализованных в геноме пластид, поскольку они непосредственно связаны с формированием

альбиносного фенотипа. Поскольку пластиды участвуют в синтезе многих растительных пигментов, представляет также интерес связь экспрессии генов пластид с биосинтезом меланина. Гены пластид были подразделены на три функциональных категории: (1) кодирующие белки, связанные с фотосинтезом; (2) кодирующие рибосомные белки; (3) прочие гены. Для каждой из категорий было оценено среднее значение и среднеквадратичное отклонение изменения экспрессии всех входящих в неё генов. С помощью теста Манна-Уитни была оценена значимость различий в изменениях уровней экспрессии для разных категорий генов.

Отдельно были рассмотрены уровни экспрессии генов, находящихся в районах Alm и Blp, в экспериментах с линией i:Bw*Alm* и i:Bw*Blp*, соответственно.

Сборка транскриптома *de novo*. В работе предложен подход к *de novo* реконструкции транскриптома путём использования нескольких программ-сборщиков и объединения результатов их работы в одну общую мета-сборку транскриптома. Сборка транскриптома контрастных по окраске органов ячменя *H. vulgare* почти изогенных линий проведена четырьмя методами: Trans-ABySS (вер.2.0.1) и Spades (вер.3.12.0), Trinity (вер.2.2.0) в режиме *de novo* и в режиме Genome-guided. *de novo* реконструкции транскриптомов линий i:Bw*Blp* и i:Bw*Alm* были проведены по отдельности.

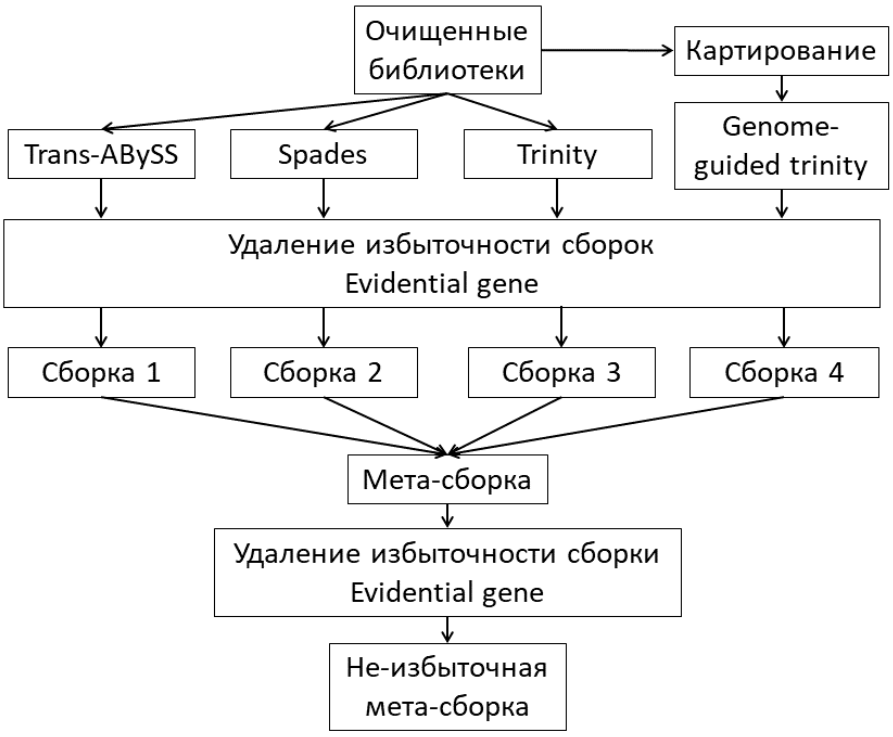

Рисунок 2. Схема *de novo* реконструкции транскриптома контрастных по окраске органов ячменя. Последовательность обработки данных на этапе независимой сборки четырьмя методами, удаления избыточности и объединения транскриптов в общую мета-сборку показана стрелками.

При запуске сборщиков Trinity и Spades на вход программ подавались все шесть библиотек, относящихся к эксперименту. Сборка программой Trans-ABySS была проведена для каждой из библиотек по отдельности с разными значениями длины *k*-меров (32, 48 и 64); полученные сборки объединялись

программой transabyss-merge. Для удаления избыточности сборок затем была использована программа tr2aacds.pl из пакета Evidential Gene (вер. от 07.05.2018). Каждая из четырех сборок была обработана этой программой по отдельности. Мета-сборка была получена путем объединения последовательностей, полученных в результате четырех независимых методов и обработки программой tr2aacds.pl. В результате, для обоих транскриптомных экспериментов были получены без-избыточные мета-сборки последовательностей транскриптов.

Анализ сборки транскриптома. Для оценки качества все отдельные сборки прошли обработку следующими программами: BUSCO (вер.3.0.2); Transrate (вер.1.0.3). Контиги мета-сборок были выровнены на геном с помощью программы rnaQUAST (вер.1.51). В результате были выделены контиги, не имеющие значимой гомологии с референсным геномом ячменя сорта Morex (называемые далее «новые контиги»). Новые контиги были также выровнены на последовательность генома ячменя сорта Bowman; те контиги, для которых гомологии с геномом этого сорта обнаружено не было, были рассмотрены более подробно. Аминокислотные последовательности открытых рамок считывания (ОРС) получены программой EvidentialGene. Для определения возможных функций и тасономической принадлежности пептидные последовательности новых контигов были выравнены с последовательностями БД NCBI nr с помощью онлайн-сервиса pblast (blastplus вер.2.13.0); для них определялся наилучший гомолог, на основе которого эти последовательности были отнесены к различным таксонам.

Глава 3. Основные результаты

Анализ транскриптора линии ячменя i:Bw*Alm*. Были использованы 6 библиотек одиночных прочтений, полученных на платформе IonTorrent PGM. Библиотеки содержат суммарно 28,5 млн. прочтений (4,7 млрд. нуклеотидов), в среднем по 4,75 млн. прочтений и 78,1 млн нуклеотидов на библиотеку.

Фильтрацию прошли 87,7% прочтений. В результате работы программы Dart выравнено в среднем 98,7% прочтений, STAR – 75,1% прочтений, Hisat2 – 61,4% прочтений, TopHat2 – 33% прочтений. После удаления генов с экспрессией ниже заданного порога и определения ДЭГ для каждого из 36 конвейеров были оценены параметры F_m, F_u, r и S_r, и проведено их ранжирование. На рис.3 приведён результат анализа производительности 36 использованных конвейеров методом главных компонент.

Наибольшую сумму рангов в эксперименте с линией i:Bw*Alm* получил конвейер, состоящий из удаления фрагментов рРНК программой Bowtie, картирования программой Dart и поиска ДЭГ пакетом EdgeR. При помощи этого конвейера 1365 генов были определены как достоверно имеющие дифференциальную экспрессию. 78 (5%) из этих генов имеют повышенную экспрессию в линии i:Bw*Alm* по сравнению с экспрессией в сорте Bowman, 1287 – пониженную.

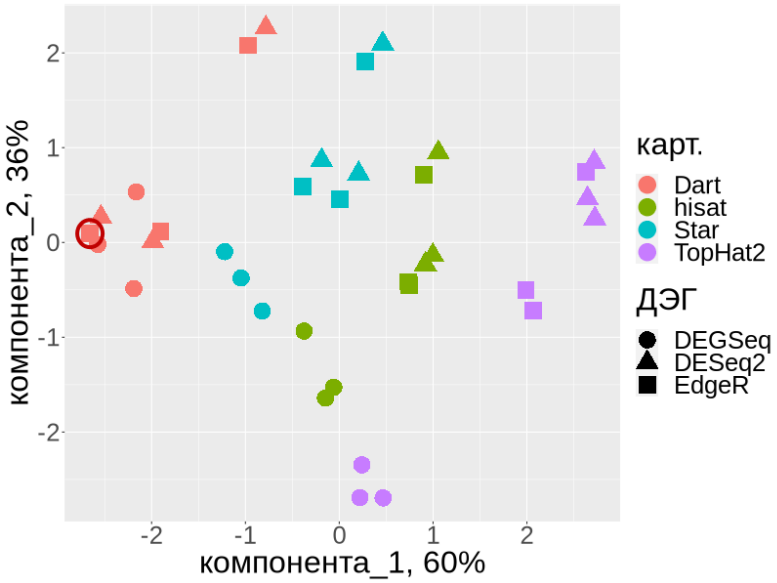

Рисунок 3. Диаграмма рассеяния результатов применения 36 конвейеров для транскриптома линии i:Bw*Alm* по первой (ось X) и второй (ось Y) главным компонентам. Красным обведён конвейер Bowtie2-Dart-EdgeR, имеющий максимальную сумму рангов среди всех конвейеров.

Функциональный анализ позволил выделить 234 термина GO как значимо обогащённые для ДЭГ с пониженной в линии i:Bw*Alm* экспрессией и 4 термина GO, значимо обогащённых для списка генов с повышенной в линии i:Bw*Alm* экспрессией. Эти термины преимущественно связаны с ответом на стресс и приведены в таблице 1.

Таблица 1. Термины генной онтологии, которыми обогащены списки генов с повышенной экспрессией в линии i:Bw*Alm*.

Термин GO	Описание	Число ДЭГ	FDR
GO:0006952	Защитный ответ	5	0,00087
GO:0006950	Ответ на стресс	8	0,004
GO:0050896	Ответ на стимул	10	0,0048
GO:0006508	Протеолиз	6	0,033

Гены с пониженной экспрессией в линии i:Bw*Alm* статистически значимо ассоциированы с 11 метаболическими путями из БД BarleyCyc (табл.2), преимущественно участвующих в фотосинтезе, фотодыхании, синтезе хлорофилла и усвоении азота, т.е. процессах, наиболее сильно нарушающихся при альбинизме растений. Для генов с повышенной экспрессией в линии i:Bw*Alm* значимо ассоциированных метаболических путей обнаружено не было. Метаболические пути, ассоциированные с ДЭГ, перечислены в таблице 2.

Дифференциальная экспрессия была обнаружена у семи генов, локализованных в районе Alm. Из них только один ген, кодирующий 40S рибосомный белок, повышает экспрессию в линии i:Bw*Alm* – его экспрессия повышена в 5 раз по сравнению с сортом Bowman ($p < 5{,}5 \cdot 10^{-4}$).

Изменения уровней экспрессии (log(FC)) для трёх функциональных групп генов, локализованных в геноме пластид, приведены на рисунке 4. Различия в изменении экспрессии между тремя группами генов пластид значимы. Так, при сравнении групп «гены фотосинтеза» и «прочие гены» значение *p*-value составило $8{,}9 \cdot 10^{-4}$, при сравнении групп «рибосомные гены» и «прочие гены» *p*-value составило 0,021, при сравнении групп «гены фотосинтеза» и «рибосомные гены» *p*-value составило $1{,}55 \cdot 10^{-5}$.

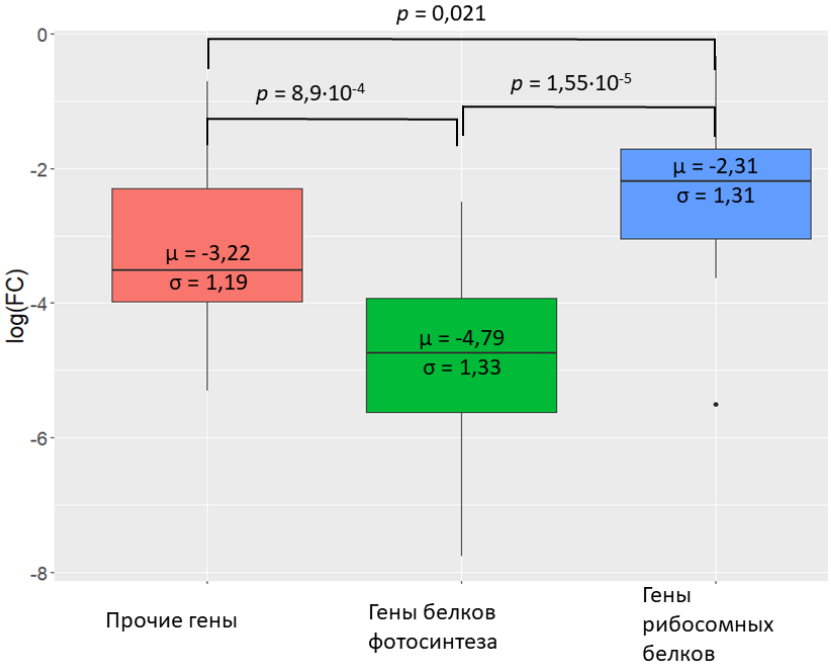

Рисунок 4. Изменение уровней экспрессии генов, входящих в три разных группы пластидных генов

Таблица 2. Метаболические пути, значимо ассоциированые с генами, понижающими экспрессию в линии i:Bw*Alm*.

Название пути в БД BarleyCyc	Число ДЭГ	FDR
Биосинтез L-глутамата	7	$5{,}34 \cdot 10^{-4}$
Фосфорилирование и дефосфорилирование NAD/NADH	38	$4{,}58 \cdot 10^{-10}$
Аэробное дыхание III (альтернативный оксидативный)	38	$1{,}12 \cdot 10^{-9}$
Биосинтез L-глутамина III	12	$1{,}9 \cdot 10^{-3}$
Шунт РБФК (Рибулозабисфосфаткарбоксилаза/оксигеназа)	28	$4{,}44 \cdot 10^{-16}$
Аэробное дыхание I (Цитохром c)	49	$1{,}10 \cdot 10^{-12}$
de novo биосинтез аденозиновых нуклеотидов	24	$2{,}13 \cdot 10^{-8}$
Цикл Кальвина-Бенсона	35	$1{,}68 \cdot 10^{-25}$
Световая фаза фотосинтеза	42	$2{,}60 \cdot 10^{-24}$
Цикл усвоения аммония II	9	$2{,}33 \cdot 10^{-5}$
Восстановление нитратов II (ассимиляционное)	7	$2{,}05 \cdot 10^{-3}$

Для поиска возможных генов-кандидатов, ассоци-ированных с мутацией в гене *Alm*, расположен-ном в ко-ротком плече хромо-сомы 3H ячменя, мы отдельно рассмотрели экспре-ссию 229 генов, лок-ализованных в этом локусе. 117 из них имеют экспрессию выше установленного порога. Из этих генов 7 понижают экспрессию в линии i:Bw*Alm*; ген 40S рибосомного белка (Ensembl ID: HORVU3Hr1G034230) повышает экспрессию в этой линии

***De novo* реконструкция транскриптома**. В результате применения методов *de novo* сборки транскриптомов линии i:Bw*Alm* и сорта Bowman итоговая мета-сборка транскриптома включала последовательности 68414 контигов. Наибольшая длина контига в сборке – 9033 нуклеотида, средняя длина – 674 нуклеотида, N50 – 940 нуклеотидов. Удаление избыточности уменьшило размер общей сборки до 62% от исходного.

Обнаруженные в контигах общей сборки ОРС кодируют 69025 белковых продуктов длинами не менее 30 аминокислотных остатков. Значения BUSCO для общей мета-сборки транскриптома и для индивидуальных сборок, из которых она была составлена, приведены на рисунке 4. Отметим, что в мета-сборке представлено большее суммарное количество последовательностей BUSCO, чем у любой из индивидуальных сборок.

Была проведена оценка экспрессии контигов общей сборки транскриптома с помощью программы kallisto. Контиги, имеющие экспрессию ниже пороговой, были удалены из рассмотрения. В результате, число транскриптов составило 55115.

В сборке транскриптома было обнаружено 943 новых контига, содержащих по одной ОРС. Из них 578 имеют гомологию к геному линии Bowman. Белковые продукты остальных новых контигов имеют гомологию с 97 известными последовательностями из базы NCBI Protein.

Последовательность транскрипта DN2647c0g1i1 отсутствует в транскриптоме сорта Bowman. В линии i:Bw*Alm* этот транскрипт имеет значимую экспрессию (TPM = 3,9). Мы предположили, что эта последовательность может являться транскриптом гена, ассоциированного с мутацией в локусе Alm. Белковый продукт ОРС транскрипта DN2647c0g1i1 демонстрирует сходство с последовательностью неаннотированного белка *H. vulgare* (NCBI id: BAK08282.1; E-value $3{,}93 \cdot 10^{-77}$) и прохибитин-1-подобным белком *Solanum pennellii* (NCBI id: XP_015060913.1; E-value $7 \cdot 10^{-20}$). В этой белковой последовательности обнаружено наличие домена SPFH-prohibitin (E-value $4{,}64 \cdot 10^{-13}$). Независимая экспериментальная проверка с помощью ПЦР на генетическом материале пшенично-ячменных дополненных линий показала локализацию этого гена в коротком плече хромосомы 3H ячменя. Известно участие прохибитин-подобных белков в стабилизации структуры митохондрий и хлоропластов; таким образом, этот белок ген может быть ассоциирован с проявлением альбинизма в линии i:Bw*Alm*.

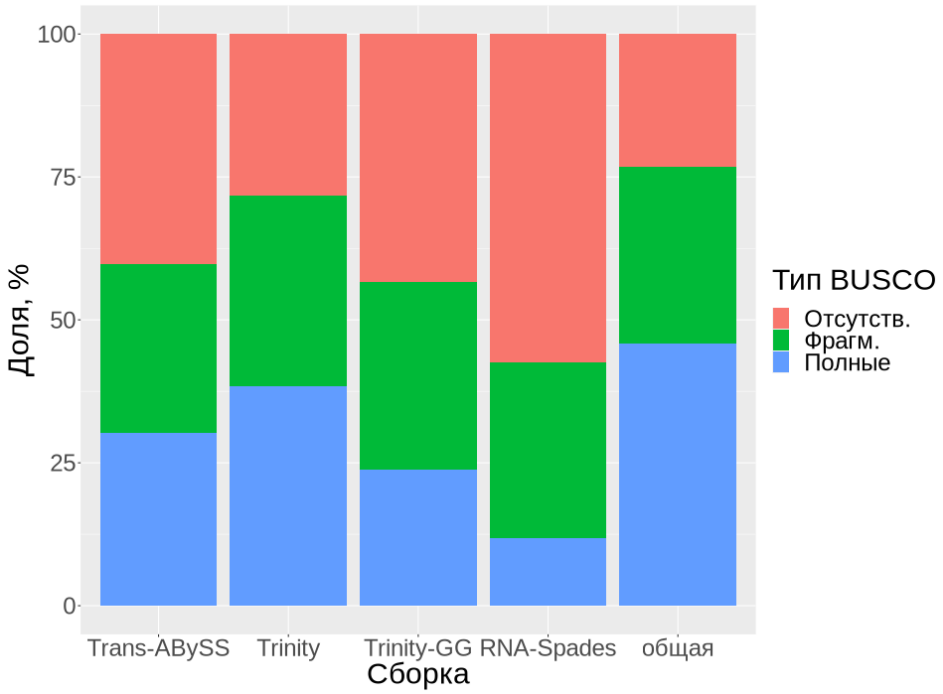

Рисунок 5. Значения BUSCO индивидуальных сборок *de novo* транскриптома и мета-сборки транскриптома.

Транскриптомный анализ линии i:Bw*Blp*

Анализ библиотек. Шесть библиотек коротких прочтений фрагментов РНК из развивающегося колоска сорта Bowman и мутантной линий i:*BwBlp* с частичным меланизмом колоса содержали в общей сложности 23128312 прочтений и 4034062730 нуклеотидов (в среднем по 3376370 прочтений). В процессе фильтрации по качеству удалено ~ 12,5% прочтений. Для поиска оптимального метода выявления ДЭГ использовали вычислительный конвейер, описанный выше. Экспериментальная оценка уровней экспрессии методом ПЦР в реальном времени была проведена для семи генов. Оптимальный конвейер включал следующие шаги: удаление рРНК путём выравнивания на последовательности рибосомных РНК, картирование на геном с помощью Hisat2, поиск ДЭГ с помощью пакета EdgeR.

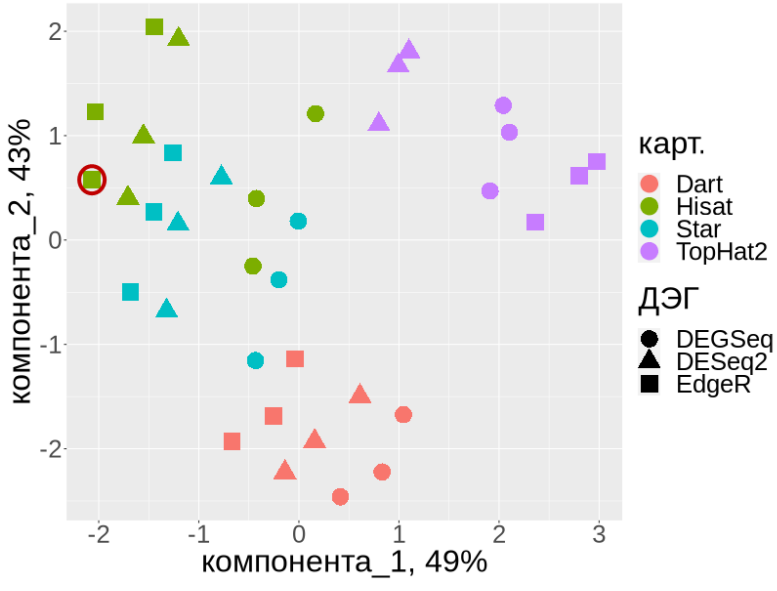

Рисунок 6. Диаграмма рассеяния результатов применения 36 конвейеров для транскриптома линии i:Bw*Blp* по первой (ось X) и второй (ось Y) главным компонентам. Красным выделен конвейер Bowtie2-HiSAT2-EdgeR, имеющий максимальную сумму рангов среди всех конвейеров и использованный для дальнейшей работы

Программа FastQC обнаружила большую представленность в образцах фрагментов рРНК злаков (родов *Triticum* и *Aegilops*). Поэтому было важно провести удаление контаминации рРНК путём картирования на последовательности некодирующих РНК. В результате было удалено в среднем 36,4% всех прочтений. Картирование с помощью метода Hisat2 оставшихся фрагментов из 6 библиотек позволило выровнять 67% прочтений. Анализ ДЭГ с помощью пакета EdgeR выявил 480 генов с пониженной и 794 гена с повышенной в линии i:Bw*Blp* по сравнению с сортом Bowman.

Эти гены аннотированы в базе данных AgriGO и для них с помощью анализа SEA было обнаружено 16 терминов генной онтологии, значимо представленных для генов с пониженной в линии i:Bw*Blp* экспрессией и 57 терминов, обогащённых для генов с повышенной в этой линии экспрессией.

Для генов с пониженной экспрессией в линии i:Bw*Blp* обогащены термины генной онтологии «Фотосинтез» (p = 0,0098), «Фотосистема I» (p = 0,034), «Тилакоид» (p = 0,013). Это говорит, что гены, участвующие в фотосинтезе, понижают свою экспрессию в этой линии. Для генов с повышенной в линии i:Bw*Blp* экспрессией были значимо обогащены термины генной онтологии «Метаболизм липидов» ($p = 2{,}1 \cdot 10^{-4}$), «Метаболизм изопреноидов» ($p = 2{,}1 \cdot 10^{-4}$), «Метаболизм ароматических аминокислот» ($p = 2{,}3 \cdot 10^{-4}$).

В базе данных BarleyCyc был обнаружен 171 метаболический путь, включающий в себя в общей сложности 152 гена с повышенной в линии i:Bw*Blp* экспрессией и 115 метаболических путей, включающих в себя 112 генов с повышенной в этой линии экспрессией. Однако, из них только 3 статистически значимо обогащены ДЭГ. Они приведены в таблице 3.

Только один ген, локализованный в районе Blp, значимо изменяет свою экспрессию. Этот ген кодирует фосфатазу пурпурной кислоты. Он повышает уровень экспрессии в линии i:Bw*Blp* в 39,8 раз по сравнению с сортом Bowman ($p = 1{,}6 \cdot 10^{-4}$).

Пластидные гены, относящиеся к разным функциональным группам, имеют разные уровни изменения экспрессии. На рис. 5 приведены средние значения изменения уровней экспрессии и их среднеквадратичные отклонения для разных функциональных групп генов пластидного генома.

Таблица 3. Метаболические пути, значимо обогащённые генами с пониженной экспрессией в линии i:Bw*Blp*.

Путь	ДЭГ	Генов	FDR
Шунт РБФК	10	63	$2{,}43 \cdot 10^{-3}$
Цикл усвоения аммония II	6	16	$9{,}56 \cdot 10^{-04}$
Цикл Кальвина-Бенсона	17	61	$8{,}82 \cdot 10^{-11}$

Достоверность различий в изменении экспрессии между группами «гены фотосинтеза» и «прочие гены» – 0,25, между группами «гены рибосомных белков» и «прочие гены» – 0,02; между группами «гены фотосинтеза» и «гены рибосомных белков» – 4,38 10^{-4}.

Такие различия в изменении уровней экспрессии указывают на перестройки в физиологической ак-тивности пластид, в частности, фотосинтеза. Это дополнительно под-тверждается участием генов, понижающих экспрессию в линии i:Bw*Blp*, в цикле Кальвина-Бенсона, и связанными с фотосинтезом терминами генной онтологии, обогащёнными для списка генов с пониженной в этой линии экспрессией.

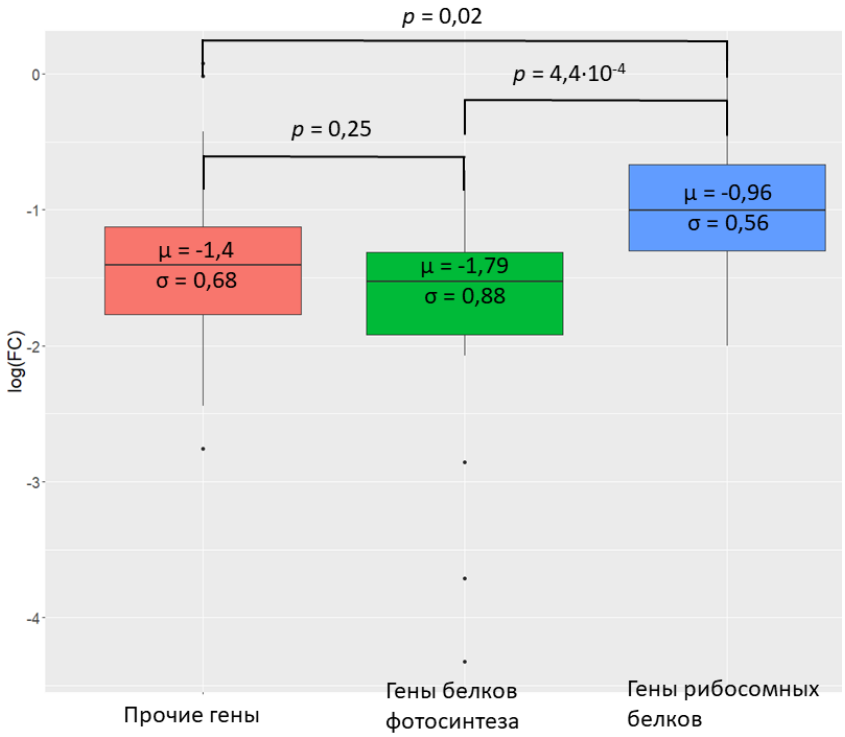

Рисунок 7. Изменение уровней экспрессии генов, входящих в три разных группы пластидных генов

Сборка транскритома *de novo*. Сборка транскриптома проводилась согласно процедуре, описанной в разделе «Методы». Полнота мета-сборки транскриптома превышает полноту отдельных сборок. 57,6% всех последовательностей BUSCO из набора для покрытосеменных организмов встречаются в не-избыточной сборке транскриптома.

После удаления контигов, имеющих экспрессию ниже пороговой, общая сборка содержит 16813 контигов. Достоверное понижение экспрессии было определено для 470 контигов, повышение – для 848 контигов.

Были обнаружены 293 новых контига, содержащих по одной ОРС. Из них 65 были удалены как артефакты сборки *de novo*, 116 – как контаминация чужеродным материалом. Из оставшихся 112 контигов 2 понижают экспрессию в этой линии, пептидный продукт одного из них имеет гомологию к белку *Brassica napus* с неизвестной функцией (CDY21223.1), уровень гомологии очень низкий (E=4,74) это позволяет отбросить данный контиг как

артефакт сборки. Пептидный продукт другого имеет высокую гомологию к цитохром-P450-подобному белку (XP_037411011.1, E=0).

Ещё 2 контига имеют повышенную экспрессию в линии i:Bw*Blp*, для их пептидных продуктов наблюдается высокая гомология к фосфатазе белков 2C 68 (XP_044946327.1, $E=4{,}14\cdot10^{-33}$) и серин/треонин протеинкиназе PBL15 (XP_044955417.1, $E=1{,}8\cdot10^{-75}$), соответственно. И протеинкиназы, и протеинфосфатазы выполняют самые разнообразные роли в растительных клетках, в том числе регулируют ответ на различные виды стресса, а также участвуют в формировании реакции на некоторые фитогормоны.

Заключение

Для биоинформатического анализа транскриптомов линий ячменя, контрастных по окраске колоса, предложен метод, основанный на выборе оптимальной комбинации программ в режиме вычислительного конвейера. Предложены критерии отбора конвейеров биоинформатической обработки данных RNA-seq, связанные с качеством картирования библиотек на референсный геном и точность определения ДЭГ. Для анализа транскриптов на основе сборки *de nov*о предложено использование мета-сборки, и продемонстрирована более высокая эффективность этого подхода.

На основе предложенных методов при анализе экспрессии генов у линии i:*BwAlm* в сравнении с контрольным сортом Bowman выявлено повышение экспрессии генов, связанных с протеолизом и ответом на стресс в развивающихся колосках ячменя при отсутствии хлорофилла. Экспрессия генов, связанных с сопутствующими фотосинтезу процессами, у линии i:Bw*Alm* снижена. Кроме того, обнаружен ген, кодирующий прохибитин-подобный белковый продукт, который экспрессируется только у линии i:Bw*Alm*, но не у Bowman.

При анализе экспрессии генов у имеющей меланиновую окраску линии i:Bw*Blp* в сравнении с Bowman установлено, что у i:*BwBlp* повышена экспрессия генов, связанных с биосинтезом жирных кислот и флавоноидов.

Выводы

1. Разработан метод биоинформатического конвейерного анализа транскриптомных данных на основе комбинации наборов компьютерных программ для оценки уровня экспрессии генов на основе как выравнивания прочтений на геном, так и сборки транскриптов *de novo*.
2. Предложен метод выбора оптимальной конфигурации биоинформатических конвейеров для анализа специфических транскриптомных данных, основанный на оценке характеристик картирования и поиска дифференциальной экспрессии генов. С помощью этого метода для двух экспериментов по сравнению транскриптомов ячменя сорта Bowman, и линий i:Bw*Alm* и i:Bw*Blp* были выбраны

оптимальные конфигурации конвейеров и выявлены дифференциально экспрессирующиеся гены.

3. В лемме ячменя линии i:Bw*Alm* большинство дифференциально экспрессирующихся генов имеют пониженный уровень экспрессии по сравнению с сортом Bowman; их функция связана с аэробным дыханием, фотодыханием и фотосинтезом, они вовлечены в метаболические пути фотосинтеза, аэробного дыхания и усвоения азота.
4. На основе анализа транскриптов, собранных *de novo* в линии i:Bw*Alm* выявлен белок-кодирующий ген, гомологичный к прохибитин-1-подобному белку *Solanum pennellii*, имеющий высокий уровень экспрессии в этой линии ячменя и нулевой в транскриптомах у сорта Bowman. Этот ген локализован в коротком плече хромосомы ячменя 3H и может быть ассоциирован с проявлением альбинизма.
5. Выявлены гены, дифференциально экспрессирующиеся в лемме ячменя линии i:Bw*Blp* и растениях сорта Bowman, функции которых связаны с биосинтезом *o*-дихинонов и фенилпропаноидов и фотосинтезом. Гены, понижающие экспрессию в линии i:Bw*Blp* вовлечены в метаболический путь ассимиляции азота, а также цикл Кальвина-Бенсона и «шунт РБФК».

Список основных работ, опубликованных по теме диссертации

1. **Шмаков Н.А.** Улучшение качества сборки *de novo* транскриптомов ячменя на основе гибридного подхода для линий с изменениями окраски колоса и стебля. Вавиловский журнал генетики и селекции. 2021; 25(1): 30-38
2. **Shmakov N.A.**, Glagoleva A.Yu., Vasiliev G.V., Afonnikov D.A., Khlestkina E.K.. Novel genomic marker for the Alm locus in barley identified based on transcriptome analysis. Current Challenges in Plant Genetics, Genomics, Bioinformatics, and Biotechnology. 2019; pp 162-164
3. Glagoleva A.Yu., **Shmakov N.A.**, Shoeva O.Yu., Vasiliev G.V., Shatskaya N.V., Börner A., Afonnikov D.A., Khlestkina E.K.. Metabolic pathways and genes identified by RNA-seq analysis of barley near-isogenic lines differing by allelic state of the Black lemma and pericarp (Blp) gene. BMC Plant Biology. 2017; 17(Suppl 1): 182
4. **Shmakov N.A.**, Vasiliev G.V., Shatskaya N.V., Doroshkov A.V., Gordeeva E.I., Afonnikov D.A., Khlestkina E.K. Identification of nuclear genes controlling chlorophyll synthesis in barley by RNA-seq. BMC Plant Biology. 2016. 16(Suppl 3): 245