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

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

Диссертация на соискание учёной степени кандидата биологических наук

Научные руководители:

к.б.н., доцент Афонников Д.А.

д.б.н., профессор РАН Хлесткина Е.К.

Новосибирск, 2023 г.

Оглавление

Список использованных сокращений

Alm – Albino Lemma

Blp – Black Lemma and Pericarp

FDR – false discovery rate

SNP – single nucleotide polymorphism, однонуклеотидные полиморфизмы

РНК – Рибонуклеиновая кислота

ДНК – Дезоксирибонуклеиновая кислота

RPO – RNA polymerase, РНК-полимераза

NEP – Nuclear-Encoded Polymerase, Полимераза, кодируемая ядерным геномом

PEP – Plastid-Encoded Polymerase, Полимераза, кодируемая пластидным геномом

TIC – Translocon Inner membrane Complex, Транслоконный комплекс внутренней мембраны

TOC – Translocon Outer membrane Complex, Транслоконный комплекс внешней мембраны

IPP – Isopenthyl diphosphate, Изопентил дифосфат

DMAPP – Dimethylallyl pyrophosphate, Диметилаллил дифосфат

MEP – Methylerythritol phosphate, Метилэритритол фосфат

Gun – Genome uncoupled, Рассогласование геномов

NADP – Nicotinamide adenine dinucleotide phosphate, Никотинамидадениндинуклеотид-фосфат

ПЦР – Полимеразная цепная реакция

ОТ-ПЦР – полимеразная цепная реакция в реальном времени с использованием обратной транскриптазы

RNA-seq – RNA sequencing, массовое высокопроизводительное секвенирование РНК

РБФК – рибулозобисфосфаткарбоксилаза/оксигеназа

BUSCO – Benchmarking universal single copy orthologues

ФГА – фосфоглицеральдегид

ФГЛ – фосфоглицерат

МТГФ – метилентетрагидрофолат

2-ОГ – 2-оксоглутарат

ГС – глутаминсинтаза

ГОГАТ – глутаматсинтетаза

Введение

Современное состояние проблемы

Растительные пигменты – собирательное название для большого количества химически разнородных соединений, придающих окраску различным органам растений. Наиболее важными растительными пигментами являются хлорофиллы – производные порфирина, участвующие в процессе фотосинтеза. В клетках сосудистых растений встречаются два типа хлорофилла, a и b, которые наиболее эффективно поглощают электромагнитное излучение с длинами волн 400-500 нм и 650-700 нм, соответственно. Спектр отраженного излучения хлорофиллов имеет максимум в области длин волн видимого света, соответствующих его зеленой части. Это и придает органам растений зелёную окраску.

Биосинтез и накопление хлорофилла у растений проходит в хлоропластах. Хлоропласты – полуавтономные органоиды, имеющие собственный геном, включающий в себя около 100-120 генов. Эти гены кодируют белки, участвующие в синтезе хлорофилла, фотосинтезе и процессах транскрипции и трансляции, а также гены тРНК и рРНК. Однако, существенная часть генов, вовлеченных в биосинтез хлорофилла, локализована в ядерном геноме растений. Число таких генов составляет по разным оценкам от 2500 до 3500 [Joyard и др., 2009; Yagi, Shiina, 2014]. Таким образом, синтез хлорофилла и фотосинтез становятся возможны только в результате тесного взаимодействия пластидного и ядерного геномов растений.

Другой тип пигментов растений, придающих окраску зерну и чешуям колоса злаков – меланины. Эти соединения окрашивают ткани и органы растений в чёрный цвет. Роль меланинов в растениях до конца не выяснена, однако показано их участие в формировании устойчивости растения к патогенам и в защите частей растения от вредителей [Glagoleva, Shoeva, Khlestkina, 2020], что, в случае накопления в оболочках зерновки ячменя, может быть сельскохозяйственно важным признаком.

Актуальность исследования

Для культурного ячменя *Hordeum vulgare* L. характерно высокое внутривидовое разнообразие по признакам окраски вегетативных и генеративных органов. Так, известным

примером частичного альбинизма ячменя фенотипически проявляется как белая окраска колосковой чешуи, цветковой чешуи (леммы), ушек и, частично, стебля и обусловлен мутацией в гене *Alm*, локализованном в коротком плече хромосомы 3H ячменя. Структура гена *Alm*, белковый продукт, его функции и конкретное воздействие, приводящее к частичной утрате хлорофилла, до сих пор неизвестны. Между тем, мутантная форма гена *Alm* представляет собой подходящую модель для понимания механизмов координированной работы ядерных и хлоропластных генов, участвующих в синтезе и распределении хлорофилла. Изучение мутантных растений с частичным альбинизмом поможет прояснить детали механизмов координированного действия ядерных и хлоропластных генов, необходимого для правильного функционирования хлоропластов [Woodson, Chory, 2008]. Мутант ячменя по гену *Alm* является одним из таких объектов. Следовательно, актуальным является установление структурно-функциональных особенностей гена *Alm*, выявление и анализ генов, дифференциально экспрессирующихся в зависимости от наличия в геноме нормального или мутантного аллеля *Alm*.

Другим примером является формирование черной окраски колосковой и цветковой чешуй и перикарпа ячменя, обусловленое аллельным вариантом гена *Blp*, который локализован в длинном плече хромосомы 1H. Белковый продукт этого гена на данный момент не известен. Чёрная окраска частей цветка в растениях этой линии вызвана содержанием в них алломеланинов – чёрных растительных пигментов полифенольной природы. Таким образом, ген *Blp* является подходящей моделью для выявления ранее неизученных или слабо изученных метаболических и генных сетей растений, к каким относится путь синтеза алломеланинов. Меланины же являются антиоксидантами, защищают семена растений от механических повреждений и поражения паразитами и патогенами. Таким образом, содержание меланинов в оболочках зерновки ячменя может быть сельскохозяйственно важным признаком.

Для решения данной задачи представляется перспективным профилирование транскриптома соответствующих сортов и линий ячменя с помощью массового высокопроизводительного секвенирования. Для этого используются платформы секвенирования второго поколения в применении к тотальной матричной РНК организма или биологического образца. Эта технология носит название RNA-seq; эксперименты такого типа генерируют большие объёмы данных, которые требуют дальнейшей

компьютерной обработки и анализа. Для этих целей были разработаны различные программные продукты.

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 данной 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 синтезом хлорофилла и фотосинтезом. 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* по сравнению с сортом Bowman, задействованы в метаболизме жирных кислот, ароматических аминокислот и изопреноидов. 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*.

Использование нескольких конвейеров биоинформатической обработки библиотек RNA-seq с последующим выбором наиболее оптимального конвейера для конкретных данных позволяет достичь большей точности в определении дифференциальной экспрессии генов. Использование нескольких сборщиков транскриптома *de novo* и последующее объединение полученных сборок в одну повышает точность и чувствительность в определении транскриптов.

Теоретическая и практическая значимость

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, 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* обнаружены транскрипты, не представленные в транскриптоме сорта Bowman. Один из таких транскриптов кодирует пептид, содержащий в своём составе домен прохибитина. Независимая эксперименитальная проверка показала, что 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*, участвуют в метаболизме жирных кислот и ароматических аминокислот. Гены, понижающие экспрессию в этой лини, участвуют в биосинтезе хлорофилла и фотосинтезе. Данные результаты позволяют предположить участие пластид в процессе синтеза меланинов в клетках леммы ячменя. Дальнейшее изучение механизмов генетического контроля синтеза меланинов является перспективным для создания сельскохозяйственно важных сортов.

Методология работы

В настоящее время широко используются методы транскриптомных исследований, основанных на массовом параллельном секвенировании матричных РНК (RNA-seq). Эти методы позволяют количественно оценить уровни экспрессии известных генов организма и реконструировать их последовательности, а также обнаружить экспрессию раннее не аннотированных генов. Важным применением метода RNA-seq является поиск значимых различий в уровнях экспрессии генов в исследуемых образцах, что в дальнейшем позволяет выдвинуть предположения о функциональной связи отдельных генов и фенотипических проявлений организмов. Кроме того, с помощью RNA-seq становится возможным обнаруживать различия в последовательностях конкретных генов у исследуемых организмов, что также может служить поводом для предположений о связи гена и признака. Таким образом, метод RNA-seq представляет собой надежный инструмент для исследования механизмов функционирования отдельных генов и генных сетей у растений.

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 *o*-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.

Личный вклад автора

Автором был проведён биоинформатический анализ библиотек коротких прочтений: фильтрация, картирование, подсчёт уровней экспрессии, поиск и функциональный анализ дифференциальной экспрессии, реконструкция транскриптома *de novo*, интерпретация полученных данных

Глава 1. Обзор литературы

1.1 Ячмень и его экономическая значимость

Ячмень (*Hordeum vulgare* L.) – вид рода ячмень (*Hordeum* L.), принадлежащий к трибе Пшеницевые (Triticeae) семейства Злаки (*Poaceae* L.). Ячмень является однолетним травянистым растением, имеющим соцветие типа колос. Плод ячменя – зерновка. Для вида ячменя *H. vulgare* характерно самоопыление.

Ячмень был одним из первых доместицированных растений, и возделывался уже около 10000 лет назад на территории Дуги Плодородия, то есть современного Ближнего Востока и Египта [Sreenivasulu, Graner, Wobus, 2008; Willcox, 2005]. Следы ячменя обнаружены в остатках пищи из неолитических поселений на этой территории, датированных началом шестого тысячелетия до нашей эры [González Carretero, Wollstonecroft, Fuller, 2017]. Дуга Плодородия считается местом одомашнивания этого растения [Mascher и др., 2016]. Однако, также высказываются предположения, что ячмень, как и многие другие культурные растения, имеет полифилетическое происхождение [Chen и др., 2012c]. Гипотеза о полифилетическом происхождении культурного ячменя высказывались ещё Н. И. Вавиловым в книге «Центры возникновения культурных растений» [Вавилов, 1926, с 84]. В пользу этой гипотезы говорят данные генетических исследований [Molina-Cano и др., 1987; Wang и др., 2016]. В качестве дополнительных центров одомашнивания ячменя называют Марокко [Molina-Cano и др., 1999], Эфиопию [Molina-Cano и др., 2005], Тибет [Chen и др., 2012a]. Указывается возможность переноса генов между разными сортами ячменя ещё около двух тысяч лет назад за счёт торговли зерном между Китаем и Ближним Востоком по Великому Шёлковому Пути [Wang и др., 2015]. Однако, другие авторы заявляют о монофилетическом происхождении культурного ячменя [Badr и др., 2000]. Таким образом, дискуссии о центрах одомашнивании этого растения ведутся до сих пор.

В наши дни ячмень – четвёртая по важности злаковая сельскохозяйственная культура, уступающая по объёмам выращивания из всех злаков только пшенице, рису и кукурузе [https://www.fao.org/faostat/en/]. В 2021 году, согласно данным FAOStat, площади, на которых во всём мире высеивался ячмень, составляли 49,5 млн гектаров

[https://www.fao.org/faostat/en/]. Крупнейшим производителем ячменя в мире является Российская Федерация [https://www.fao.org/faostat/en/].

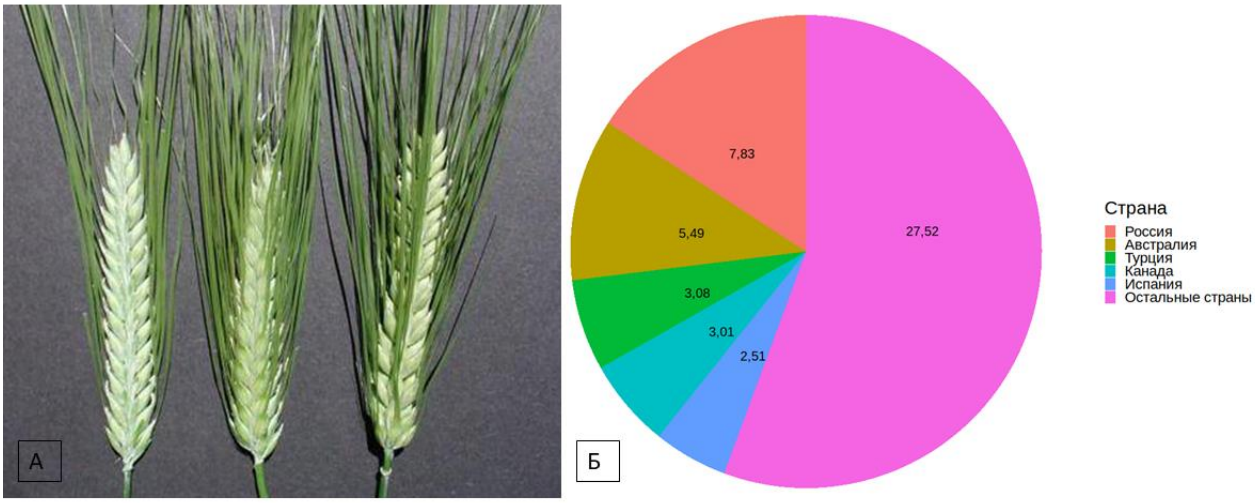

Рис. 1. (А) – колосья ячменя, слева направо: линия CIho 4196, мутант G07-014, сорт Morex; изображение взято из базы данных GrainGenes [https://wheat.pw.usda.gov/ggpages/bgn/38/bgn38cover.htm] (Б) – площади, на которых в разных странах высевался ячмень в 2021 году, млн га [https://www.fao.org/faostat/en/].

Содержание белка в зерне – важный признак качества зерновых культур, в том числе для ячменя. Зёрна ячменя с высоким содержанием белка подходят для кормовых нужд, в то время как зёрна с низким содержанием белка – для пивоварения [See и др., 2002]. В основном для пивоварения используются зёрна ячменя с содержанием белка 9,5% - 11,5% [Guo и др., 2016]. Большая часть производимого ячменя используется в кормовых целях, примерно одна треть – для производства пива и спирта, для продовольственных целей – в десять раз меньше.

Ячмень имеет большое значение как продовольственная культура для населения некоторых стран Африки и Азии. Но в последние годы возрастает интерес к ячменю как источнику питательных веществ и в странах, где ячмень ранее преимущественно рассматривался лишь как источник кормов и сырья для спиртовой и пивовареной промышленности – в Германии, Франции, США и др. странах. Возникший интерес связан с популярным трендом функционального питания. Наличие в зерновке ячменя высокого

содержания биоактивных веществ, в частности – полифенольных соединений, включило его в ряд продуктов для функционального питания [Pihlava, 2014].

Необходимо отметить, что ячмень является культурой, устойчивой ко многим неблагоприятным факторам среды, патогенам и вредителям, что даёт преимущество этой культуре для выращивания в широком диапазоне почвенных и климато-географических условий, а следовательно ставит в ряд наиболее перспективных культру для удовлетворения пищевых потребностей растущего населения планеты [Newton и др., 2011].

1.2 Геном ячменя

Ячмень, имеющий диплоидный геном (2n = 2x = 14), – удобный модельный объект для генетических исследований злаков трибы Triticeae и в частности для пшеницы мягкой, имеющей гексаплоидный геном. Результаты, полученные при исследовании ячменя, например, по выявленным генетическим механизмам устойчивости к неблагоприятным условиям окружаеющей среды, могут быть далее использованы в исследованиях пшеницы [Dawson и др., 2015].

Как и у пшеницы, хромосомы ячменя имеют достаточно большие размеры за счет высокого содержания фракции повторяющихся последовательностей (ПП), характерного для трибы Triticeae. Наименьший размер в сборке генома ячменя версии IBSC_v2 имеет хромосома 1H, 558 млн. пар оснований, наибольший – хромосома 2H, 768 млн. пар оснований. Общий размер генома оценивается как 5,5 млрд. пар оснований. Долгое время большой размер хромосом и высокая доля ПП осложняли секвенирование и сборку полного генома ячменя [Sreenivasulu, Graner, Wobus, 2008].

В 2006 году был основан проект Международной инициативы секвенирования генома Ячменя (International Barley Genome Sequencing Consortium, IBGSC). В ходе работ по этому проекту в 2012 году были проведены секвенирование и сборка генома ячменя изогенной линии Morex [IBGSC, 2012]. Для сборки были использованы 571 тысяча искусственных бактериальных хромосом. Было собрано 9265 контигов со значением N50 равным 904 тысяч пар оснований, имеющих суммарную длину 4,98 млрд. пар оснований. Также, полный геном был секвенирован методом дробовика с помощью платформы для секвенирования нового поколения Illumina GAIIx. 6437 контигов были привязаны к конкретным позициям на хромосомах ячменя. Эти контиги имеют суммарную длину 4,56

млрд пар оснований, что оценивается как 90% от всего размера генома ячменя. В этой сборке 26159 локусов были с высокой степенью достоверности аннотированы как гены, причём авторы отмечают их как гены «с высокой степенью уверенности». В то же время, ещё 53220 локуса выделены как гены «с низкой степенью уверенности».

Помимо этого, геномы сортов Barke, Morex, Igri и Bowman были секвенированы с помощью приборов Illumina GAIIx и Hiseq 2000, и геном ячменя линии Haruna Nijo был секвенирован с помощью платформы Roche 454 GSFLX Titanium. Далее, полученные библиотеки сиквенсов линий Morex, Bowman и Barke прошли фильтрацию по качеству и сборку *de novo* с помощью программы CLC Assembly Cell v3.2.2 [IBGSC, 2012]. Сборки были проведены до состояния контигов со средними длинами 700 пар оснований, 736 пар оснований и 856 пар оснований для линий Morex, Barke и Bowman соответственно. Количество контигов составляет 2,67 млн, 2,74 млн и 2 млн для линий Morex, Barke и Bowman соответственно.

Работы по секвенированию генома ячменя на этом не закончились, и через 5 лет была выпущена новая версия сборки генома [Mascher и др., 2017]. Авторы секвенировали 87 тысяч искусственных бактериальных хромосом, получив в сумме 4,5 триллиона пар оснований. Они составили из секвенированных BAC-клонов супер-скаффолды со значением длин N50 равным 1,5 млн. пар оснований. Супер-скаффолды были отнесены к конкретным позициям на физической карте хромосом с помощью POPSEQ-маркеров. В итоге была получена сборка суммарной длиной 4,95 млрд. пар оснований, из которых 4,54 точно локализованы на хромосомах. Были определены 83105 потенциальных генов, включая 39841 белок-кодирующий ген.

Ячмень является подходящей моделью для исследования генетики злаков и особенностей генетического контроля реакции растений на изменяющиеся условия среды [Dawson и др., 2015].

1.3 Биология пластид

Пластиды – полуавтономные органоиды, встречающиеся в клетках растений и некоторых простейших. Как и другой тип полуавтономных органоидов – митохондрии – пластиды имеют свой геном, который иногда также называют “пластом”. В пластидах есть собственный аппарат транскрипции генов и аппарат синтеза белка, состоящий из рибосом

прокариотического типа – 70S. Они отделены от цитоплазмы клетки-хозяина двухслойной мембраной и способны к делению [Cackett и др., 2022]. В свете всего этого, на сегодняшний день считается общепризнанным, что пластиды появились в результате эндосимбиоза ранней эукариотической клетки со свободноживущей цианобактерией. В качестве цианобактерий, наиболее близких к потенциальному свободноживущему предку пластид современных растений, называют *Gloeomargarita lithophora* [Ponce-Toledo и др., 2017], однако в целом этот вопрос на сегодняшний день остаётся открытым [Lewis, 2017].

Транскрипционный аппарат пластид представлен двумя типами РНК-полимераз. Первый тип – РНК-полимераза бактериального типа, состоящая из пяти субъединиц, в работе которой участвуют сигма-факторы [Pfannschmidt и др., 2015]. Субъединицы этой полимеразы кодированы генами, локализованными в пластоме, в то время как гены, кодирующие сигма-факторы, локализованы в ядерном геноме клетки-хозяина. Другой тип – односубъединичные РНК-полимеразы фагового типа, которые, в свою очередь, подразделяются на типы RpoTp, RpoTmp [Emanuel и др., 2004]. Первый из этих белков после синтеза на рибосомах в цитоплазме клетки направляется в пластиды, второй может быть направлен как в пластиды, так и в митохондрии [Hedtke, Börner, Weihe, 2000]. Существует также белок RpoTm, который схож по строению с первыми двумя, но направляется исключительно в митохондрии. Гены, кодирующие все три этих белка, локализованы в ядерном геноме растительной клетки [Demarsy и др., 2006]. Пластидные РНК-полимеразы были названы по месту локализации генов, кодирующих их белки – кодируемая ядерным геномом (Nuclear-Encoded Polymerase, NEP) и кодируемая пластидным геномом (Plastid-Encoded Polymerase, PEP) РНК-полимеразы, соответственно.

После открытия двойственной природы транкрипционного аппарата пластид изначально было высказано предположение, что эти типы полимераз различаются по своим функциям [Hajdukiewicz, Allison, Maliga, 1997]. Предполагалось, что NEP транскрибирует гены, связанные с транскрипционным и трансляционным аппаратами пластид, соответственно – гены PEP, гены рибосомных белков и гены транспортных РНК пластид. PEP же, как считалось, транскрибирует гены, связанные с синтезом хлорофилла и фотосинтезом. В соответствии с этим, гены пластома подразделяли на три категории - транскрибируемые исключительно PEP, транскрибируемые исключительно NEP, и гены,

способные транскрибироваться и той, и другой полимеразами [Hajdukiewicz, Allison, Maliga, 1997].

Однако в дальнейшем были получены свидетельства, опровергающие эту точку зрения. Так, в пропластидах в клетках сухих семян растений уже присутствуют транскрипты генов, кодирующих белки PEP, и происходит транскрипция некоторых пластомных генов [Emanuel и др., 2004].

Несмотря на то, что пластиды имеют собственный геном, он существенно редуцирован, и содержит в большинстве случаев около ста генов [Börner и др., 2015]. При этом, протеом пластид содержит, как правило, около 3-4 тысяч белков [Zoschke, Bock, 2018]. Большинство из этих белков кодированы генами, локализованными в ядерном геноме [Khan, Lindquist, Aronsson, 2013], и после синтеза на рибосомах в цитоплазме клетки доставляются в пластиды. Основной механизм транспорта белков в пластиды – белковые комплексы, называемые “Транслоконные комплексы внутренней и внешней мембран” (Translocon Inner-Outer membrane Complex, TIC и TOC, соответственно) [Richardson, Singhal, Schnell, 2017]. Белки комплекса TOC опознают короткие сигнальные последовательности, встречающиеся у белков, направляемых в пластиды, и перемещают их в межмембранное пространство, откуда белок, при наличии у него другой сигнальной последовательности, перемещается в строму с помощью комплекса TIC [Sadali и др., 2019].

Последовательность пластома, входящие в него гены и их относительное расположение являются эволюционно консервативными [Amiryousefi, Hyvönen, Poczai, 2018]. Более того, изучение нефотосинтезирующих паразитических растений, геном которого в силу их паразитизма оказывается существенно редуцированным, показывает, что их пластом сохраняет свою структуру, хотя и может быть редуцирован ещё сильнее – пластом в клетках *Epifagus virginiana* содержит 42 гена, и не имеет генов, связанных с фотосинтезом и фотодыханием [Ems и др., 1995]. В геноме цветкового растения *Rafflesia lagascae* Blanco не была обнаружена последовательность пластидных хромосом [Molina и др., 2014]. Однако, что несмотря на это, пластиды в клетках раффлезии по-прежнему сохраняются.

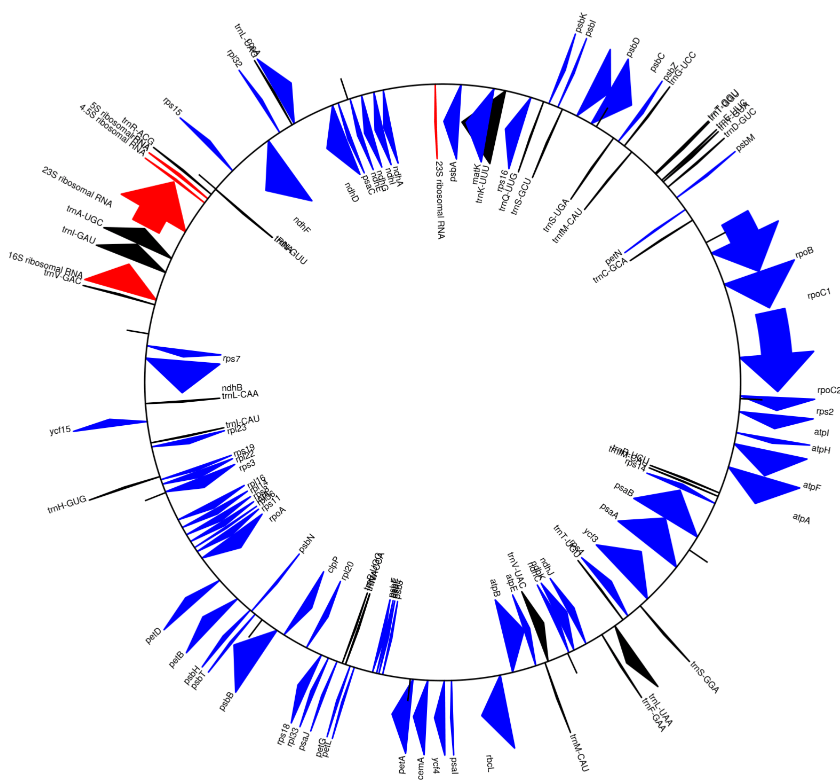

Рис. 2. Пластом ячменя. Красным цветом отмечены гены рибосомных РНК, чёрным – гены транспортных РНК, синим – белок-кодирующие гены.

Пластиды имеют несколько устойчивых форм, наиболее известная из которых – хлоропласты. В хлоропластах содержится хлорофилл, который придаёт клеткам растений зелёную окраску и даёт им способность к фотосинтезу. Однако, помимо хлоропластов, известно множество других устойчивых форм пластид: пропластиды, этиопласты, хромопласты, лейкопласты, геронтопласты и другие.

В клетках зародышей растений пластиды представлены в качестве пропластид. Они содержат везикулы, но не тилакоиды [Pogson, Ganguly, Albrecht-Borth, 2015]. По мере развития пластиды, проламеллярное тело трансформируется сначала в основную ламеллу, из которой потом формируются мелкие граны, которые далее преобразуются в полноразмерные граны. Затем, в фазе созревания зерна, они на короткое время трансформируются в хлоропласты, в результате чего семена способны к фотосинтезу и усвоению углекислого газа. Далее, в фазе высыхания, хлоропласты дедифференцируются и

трансформируются в эопласты, бесцветные и не способные к фотосинтезу, и остаются в этой форме до прорастания семян [Allorent и др., 2013]. После этого эопласты в клетках побега растения трансформируются снова в хлоропласты, способные к фотосинтезу, в то время как в клетках гипокотиля и корня они трансформируются в амилопласты [Demarsy и др., 2012].

Множество молекул направляется из цитоплазмы в пластиды, некоторые из них способны влиять на экспрессию генов пластид. Однако, при этом также и из пластид в цитоплазму поступает целый ряд соединений, которые влияют на экспрессию ядерных генов. Эти вещества называют ретроградными сигналами. Известно, что в качестве ретроградных сигналов выступают такие соединения, как Mg-протопорфирин IX. Выделяют ретроградные сигналы, связанные с:

1) метаболитами из пути синтеза тетрапирролов;

2) экспрессией пластидных генов;

3) активными формами кислорода;

4) нарушение метаболизма пластид.

В основном это связано с ответами на стресс и контролем деятельности пластид [Hernández-Verdeja, Strand, 2018].

Многие мутанты, в которых нарушено развитие пластид, гибнут ещё на стадии эмбриона [Shi, Theg, 2013]. Летальность этих мутаций на стадии эмбриона говорит о том, что к гибели организма приводит не энергетический голод из-за неспособности фотосинтезировать, а некие принципиальные нарушения развития клеток, связанные с нарушениями развития пластид [Pogson, Ganguly, Albrecht-Borth, 2015]. Выделяют три типа локализованных в пластидах белков, мутации в генах которых связаны с эмбриональной летальностью у *Arabidopsis thaliana* (L.) Heynh. – ферменты, необходимые для синтеза аминокислот, нуклеотидов или жирных кислот; белки, участвующие в транспорте и модификации белков в пластидах; белки, участвующие в трансляции в пластидах [Bryant и др., 2011].

Таким образом, пластиды оказывают огромное влияние на растительную клетку, на растительный организм в целом и даже на экосистемы, выделяя летучие соединения, которые улавливаются другими растениями и могут влиять на их метаболизм [Bobik, Burch-Smith, 2015]. Возможное объяснение этого – то, что пластиды выступают в роли центра

чувствительности растительной клетки к окружающим условиям [Chan и др., 2016]. В меняющихся условиях среды растения должны сохранять способность к фотосинтезу, поэтому с эволюционной точки зрения оправдано, что пластиды, где непосредственно происходит фотосинтез, и являются сенсорами факторов окружающей среды [Nikkanen, Rintamäki, 2019]. Однако природа обратных сигналов, поступающих из пластид в ядро и влияющих на экспрессию ядерных генов, до сих пор до конца не изучена [Liebers и др., 2017]. В то же время, это является принципиальным вопросом биологии растений [Xiao и др., 2012].

1.4 Альбинизм растений

В процесс фотосинтеза вовлечено огромное количество белков. Существуют данные, что у *A. thaliana* до трети всех генов меняют свою экспрессию в ответ на изменение условий освещения [Chen и др., 2010]. Большая часть этих белков кодированы генами, локализованными в ядерном геноме, однако некоторые закодированы в генах пластома. В результате, для достижения стехиометрического равновесия белков, участвующих в одной и той же реакции, необходима точная координация экспрессии пластидных и ядерных генов.

Помимо генов, участвующих в фотосинтезе, большое их количество вовлечено в процесс биосинтеза хлорофиллов, который происходит в пластидах [Brzezowski, Richter, Grimm, 2015]. Первые стадии биосинтеза хлорофилла состоят в синтезе протопорфирина IX из 5-аминолевулиновой кислоты. Гены, кодирующие ферменты, вовлечённые в этот процесс, локализованы в ядерном геноме. Далее, фермент Mg-хелатаза (EC 6.6.1.1) катализирует присоединение иона магния Mg^{2+} к молекуле протопорфирина IX, в результате чего образуется Mg-протопорфирин IX [Brzezowski, Richter, Grimm, 2015]. В то же время, фермент протопорфирин/копропорфирин феррохелатаза (4.99.1.1, коротко Fe-хелатаза) катализирует присоединение иона железа Fe^{2+} к молекуле протопорфирина IX с образованием гема [Woodson, Perez-Ruiz, Chory, 2011]. Эти процессы происходят в пластидах. Таким образом, процесс синтеза хлорофилла в растительных клетках сопряжён с синтезом гема, при этом фермент, катализирующий образование гема – Fe-хелатаза – конкурирует за субстрат с Mg-хелатазой, которая катализирует важную стадию метаболического пути образования хлорофилла.

Далее, Mg-протопорфирин IX в результате нескольких реакций преобразуется в хлорофиллид *a*, который с помощью фермента хлорофилл *a*-синтазы (EC 2.5.1.62) этерифицируется с фитол дифосфатом, который далее восстанавливается по трём двойным связям, и так образуется хлорофилл *a*. Фитол дифосфат образуется в клетках растений в результате работы ферментов метаболического пути биосинтеза метилэритритол фосфата [Rodríguez-Concepción, Boronat, 2002]. Этот метаболический путь, локализованный в пластидах, начинается с того, что в ходе реакции, катализируемой деоксиксилозофосфатсинтазой (EC 4.1.3.37), глицеральдегид-3-фосфат взаимодействует с пируватом с образованием деоксиксилозофосфата и выделением углекислого газа. Далее, после череды реакций, образуются изопентил дифосфат (IPP) и его изомер диметилаллил дифосфат (DMAPP), взаимные превращения которых катализируются ферментом изопентилдифосфатизомеразой (EC 5.3.3.2). В результате действия геранилгеранилдифосфатсинтазы (EC 2.5.1.29) из этих соединений образуется фитол дифосфат, который, как было сказано ранее, вступает в реакцию с хлорофиллидом *a*, в итоге образуя хлорофилл *a*.

Работа метаболических путей синтеза тетрапирролов и MEP должна быть точно координирована, поскольку нарушение соотношения продуктов этих путей в клетках может привести к накоплению промежуточных продуктов синтеза хлорофилла, таких, как протопорфирин IX или протохлорофиллид *a*, некоторые из которых токсичны [Rodríguez и др., 2013]. Обработка растений фосфомицидином, который блокирует работу пути синтеза MEP, приводит к летальности на свету, когда предшественники хлорофилла могут синтезироваться, но не в темноте, когда синтез предшественников хлорофилла не происходит, и не приводит к летальности в том случае, если путь синтеза тетрапирролов тоже блокированы [Kim и др., 2013b]. Отметим, что протохлорофиллид оксидоредуктаза и некоторые другие ферменты из пути биосинтеза хлорофилла функционируют только на свету [Wu и др., 2018].

Механизмы развития пластид, синтеза хлорофилла и хлорофилл-связывающих белков, как и других белков, принимающих участие в фотосинтезе, имеют огромное значениие для жизни не только растений, но и всей биосферы. Поэтому данные механизмы в течение уже долгого времени были предметом пристального внимания научного сообщества. Для исследования этих механизмов используются мутанты с нарушением

синтеза хлорофилла, имеющие альбиносный, частично альбиносный, бледно-зелёный и желтоватый фенотипы. У различных растений описаны разнообразные гены, мутации в которых связаны с нарушениями синтеза хлорофилла и приводят к образованию альбиносного, желтоватого или бледно-зелёного фенотипов [Bae и др., 2001; Chen и др., 2009a; Tang и др., 2018; Toshoji и др., 2012; Zhang и др., 2018]. Помимо генов, кодирующих белки, вовлеченные в путь биосинтеза хлорофилла, выделяют и другие – гены, кодирующие белки пластидных рибосом [Qiu и др., 2018], пентатрикопептиды [Liu и др., 2018], ферменты из пути биосинтеза пуриновых нуклеотидов [Cao и др., 2019] и другие.

Популярным объектом такого рода исследований является *A. thaliana*. Было исследовано множество мутантов этого растения, имеющих фенотип, так или иначе связанных с нарушением синтеза хлорофилла [Casanova-Sáez и др., 2014; Martínez-Zapater, 1993; Sakamoto, 2003; Waters и др., 2006]. Значительной вехой в исследовании синтеза хлорофилла, развития пластид и влияния их сигналов на работу ядерного генома стало открытие в 1993 году у арабидопсис трёх мутантов, у которых экспрессия ядерных генов *Cab* и *Rbsc* не зависит от стадии развития пластид [Susek, Ausubel, Chory, 1993]. Фенотип, наблюдаемый у этих мутантов, был назван "рассогласование геномов" (Genomes uncoupled, сокращённо 'gun'). Как было установлено впоследствии, многие другие ядерные гены, кодирующие связанные с фотосинтезом белки, также не изменяют свою экспрессию в растениях с gun-фенотипом, если в клетках этих растений нарушается развитие пластид и синтез хлорофиллов [Garnik и др., 2019]. Гены, мутации в которых приводили к формированию такого фенотипа, по аналогии с этим фенотипом были названы *gun1*, *gun2* и *gun3*.

Впоследствие были обнаружены другие мутанты, имеющие фенотип gun [Larkin, Brown, Schiefelbein, 2003; Mochizuki и др., 2001; Woodson, Perez-Ruiz, Chory, 2011]. На сегодняшний день описано шесть мутантов, имеющих фенотип gun. Дальнейшие работы позволили охарактеризовать гены *gun* и привязать их белковые продукты к конкретным биологическим функциям. Ген *gun1* кодирует локализованный в пластидах пентатрикопептидный белок [Koussevitzky и др., 2007]. Позднее было установлено, что этот белок участвует в регуляции биосинтеза тетрапирролов [Shimizu и др., 2019]. Остальные пять генов *gun* кодируют ферменты, действующие на разных стадиях метаболического пути биосинтеза тетрапиррольных соединений. *gun2* кодирует гем-оксигеназу 1, *gun3* -

фитохромобилин синтазу, и *gun6* - пластидную феррохелатазу [Chi, Sun, Zhang, 2013], то есть ферменты, участвующие в ветви синтеза гема общего пути биосинтеза тетрапирролов. Гены *gun4* и *gun5* кодируют, соответственно, D- и H-субъединицы магний-хелатазы [Larkin, 2016], то есть первого фермента в ветви синтеза хлорофилла общего пути биосинтеза тетрапиррольных соединений.

Арабидопсис – не единственный объект, использованный для исследования регуляции развития пластид и экспрессии связанных с этим процессом генов. Для этой цели используют множество других растений, в том числе злаки. Работы проводятся на кукурузе, как хорошо изученном объекте [Rodríguez и др., 2013]; на рисе, как на высоко востребованном объекте [Cao и др., 2019; Liu и др., 2018; Qiu и др., 2018]. Но также достаточно большое количество исследований использует в качестве биологического объекта ячмень [Hess и др., 1994; Landau и др., 2007; Prina, 1996; Prina и др., 2003].

Именно на ячмене было проведено исследование, которое позволило выдвинуть гипотезу о влиянии экспрессии генов пластид на экспрессию генов ядра [Bradbeer и др., 1979]. Авторы наблюдали уменьшение количества ферментов фосфорибулокиназы и D-глицеральдегид-3-фосфат NADP+ оксидоредуктазы в клетках белых листьев растений ячменя мутантных линий Saskatoon и albostrians. Оба этих фермента кодированы ядерными генами, их синтез происходит на цитоплазматических рибосомах, локализованы эти ферменты в пластидах. В белых листьях двух этих мутантных линий ячменя пластиды не содержат хлорофилла, и их развитие происходит аномально. Из этого авторы сделали вывод, что состояние пластид может влиять на экспрессию ядерных генов, что противоречило существовавшей в то время парадигме, согласно которой только работа ядерных генов могла влиять на работу генов полуавтономных органоидов, но не наоборот [Ellis, 1977]. Эта работа положила начало изучению такого явления, как обратные сигналы из пластид в ядро [Börner, 2017].

Мутанты, имеющие описанный выше фенотип gun, были также обнаружены у ячменя [Gadjieva и др., 2005]. Помимо этого, были исследованы другие мутанты ячменя, имеющие частично альбиносный, бледно-зелёный или желтоватый фенотипы. [Hagemann, Scholz, 1962; Møller и др., 1997; Qin и др., 2015; Svensson и др., 2006].

Всё это говорит о том, что ячмень – перспективный объект для дальнейшего изучения развития пластид, синтеза хлорофилла, обратных сигналов из пластид в ядро, и

генетической регуляции всех этих процессов. Также отметим, что во многих работах, направленных на изучение развития пластид, синтеза хлорофилла и регуляции и нарушений этих процессов, проводится профилирование транскриптома растения с целью изучения экспрессии генов, связанных с фотосинтезом и синтезом хлорофилла, а также остальных генов. Для этих целей в разных работах используются как микрочипы [Grübler и др., 2017; Rodríguez и др., 2013], так и массовое параллельное секвенирование транскриптома с помощью секвенирования нового поколения, то есть RNA-seq [Gang и др., 2019; Nguyen и др., 2014]. В том числе эксперименты RNA-seq проводятся и на ячмене в качестве объекта [Bian и др., 2019; Tan и др., 2019].

1.5 Меланизм растений

Растительные пигменты включают большой набор соединений, создающих окраску отдельных органов растений. В то время как основная функция хлорофилла, наиболее распространённого растительного пигмента на Земле, состоит в улавливании энергии фотонов, что делает возможным процесс фотосинтеза, другие пигменты выполняют множество разнообразных функций. Разные типы окраски органов и тканей в растениях создаются различными химическими соединениями. Один из наиболее широко представленных классов растительных пигментов – флавоноиды [Mierziak, Kostyn, Kulma, 2014]. С химической точки зрения флавоноиды представляют собой производные флавона, то есть гетероциклические соединения полифенольной природы. Известно более девяти тысяч флавоноидных соединений [Buer, Imin, Djordjevic, 2010]. Эти пигменты придают органам растений различные типы окраски: фиолетовую, синюю, коричневую и другие [Zhu, 2018]. Помимо создания окраски, флавоноиды также выполняют другие функции в растениях, такие как защита от патогенов [Wang и др., 2022] и абиотических стрессовых факторов [Ghitti и др., 2022], регуляция транспорта ауксина [Peer, Murphy, 2007] и прочие.

Помимо флавоноидов, тёмную окраску органам растений могут придавать меланины. Меланины – обширный класс химических соединений – производных фенолов [Tarangini, Mishra, 2014]. Меланины встречаются у бактерий, грибов, растений и животных [Kim, Uyama, 2005]. Меланины, встречающиеся в растениях, называют «алломеланины», тогда как у других таксономических групп встречаются другие типы меланинов – феомеланины и нейромелалины [Charkoudian, Franz, 2006] у животных, эумеланины у

животных, бактерий и грибов. При этом, алломеланины также встречаются у грибов и бактерий.

Алломеланины, механизм их синтеза и его регуляция у растений изучены хуже, чем биохимия и генетика синтеза других типов меланина у прочих таксономических групп. Изучению алломеланинов было уделено меньше внимания, чем изучению других растительных пигментов [Varga и др., 2016], поскольку алломеланины, как правило, представлены в растениях одновременно с другими пигментами, и функционал алломеланинов долгое время был не ясен. В целом же, химические исследования этих веществ затруднены низкой растворимостью меланинов в большинстве широко используемых растворителей (Park 2007), что осложняет их выделение; однако, меланины хорошо растворимы в щелочных растворах [Kamei и др., 1997]. Наконец, молекулы алломеланинов, в отличие от других типов меланина, не содержат атомов азота [Solano, 2014]. По этой причине алломеланины ранее не относили к числу меланинов. Считалось, что меланины животных представляют гораздо больший биологический интерес, чем меланины прочих групп организмов [Prota, 1980]. В свете всего перечисленного видно, что алломеланины до недавнего времени не привлекали внимания исследователей.

Как следствие, биохимия меланинов и пути их биосинтеза были лучше изучены у грибов [Butler, Gardiner, Day, 2009; Eisenman, Casadevall, 2012; Pal, Gajjar, Vasavada, 2014], бактерий [Hernández-Romero, Solano, Sanchez-Amat, 2005; Kelley и др., 1990], а также у животных [Bourgeois и др., 2016; Galván, Solano, 2016], в том числе у человека [Itou, Ito, Wakamatsu, 2019; Sitiwin и др., 2019]. Синтез всех известных типов меланинов, за исключением алломеланинов, начинается с единственного предшественника – тирозина [Cao и др., 2021; D'Alba, Shawkey, 2019]. Однако, есть сообщения о наблюдении синтеза меланина с использованием триптофана в качестве предшественника в бактерии *Rubrivivax benzoatilyticus* JA2 [Ahmad и др., 2020].

Отличительной же чертой алломеланинов является то, что у этих соединений нет единственного предшественника – в качестве предшественников разных алломеланинов выступают катехол, хиноны и 1,8-дигидроксинафтален [Cao и др., 2021; D'Alba, Shawkey, 2019]. Синтез алломеланинов осуществляется полифенолоксидазами.

В настоящий момент изучение алломеланинов, процесса их биосинтеза и механизмов генетической регуляции кажется перспективным направлением. Появляются

данные о функциях алломелалинов и их важности для выживания растений. Так, алломеланины, накапливающиеся в оболочке семян подсолнечника, предохраняют их от поражения личинками насекомых [Pandey, Dhakal, 2001]. У арбуза *Citrullus lanatus* Matsum. & Nakai успешно проросли 86,5% семян из числа тех, в оболочке которых содержится большое количество меланинов, что придаёт им чёрную окраску, в то время как среди семян со светлой оболочкой и малым содержанием меланина только 37,5% успешно проросли [Mavi, 2010]. Меланины, экстрагированные из семян арбуза, показывают антиоксидантные и антибактериальные свойства, и задерживают жёсткое излучение [Łopusiewicz, 2018]. Таким образом, меланины имеют как важность для выживания и успешного размножения растения, так и практическое применение в разных видах промышленности.

У животных меланины синтезируются и накапливаются в особых клетках – меланоцитах. Меланоциты – дендритные клетки нейроэктодермы [D'Mello и др., 2016]. У растений же специализированных клеток, содержащих меланины, нет. Ранее считалось, что синтез и накопление алломеланинов происходит в межклеточном пространстве. Однако, последние данные показывают, что внутриклеточные области флуоресценции хлорофилла и меланина в клетках перикарпа у растений ячменя линии, отличающейся частичным меланизмом колоса, совпадают [Shoeva и др., 2020]. Это, вкупе с локализацией большого количества растительных полифенолоксидаз в пластидах [Boeckx и др., 2015] и присутствии в пластидах большого количества фенольных соединений, являющихся предшественниками меланина и субстратом для полифенолоксидаз [Boeckx и др., 2017a], позволяет предположить, что у ячменя меланин синтезируется и накапливается в пластидах. В связи с этим был предложен новый термин, характеризующий пластиды, в которых происходит накопление меланинов – меланопласты [Shoeva и др., 2020]. На данный момент не установлено, у всех ли растений накопление меланинов происходит в пластидах, или же этот процесс уникален для ячменя [Shoeva и др., 2020]. В целом, пластидная локализация алломелалнинов и белков, участвующих в пути их биосинтеза, позволяет предположить, что синтез меланинов в растениях требует точной координации между пластидами и ядром. В таком случае, изучение этого процесса может пролить свет на генетические механизмы связи двух геномов – пластидного и ядерного – и типы сигналов, которыми они обмениваются.

1.6 Методы транскриптомных исследований

Понятие «транскриптом» было введено в 1997 году и было определено как набор всех генов, экспрессирующихся в данной популяции клеток, включая последовательность и количественный уровень экспрессии каждого из этих генов [Velculescu и др., 1997]. Позднее было также сформулировано следующее определение транскриптома – совокупность всех транскриптов, присутствующих в клетке на определённой стадии развития или в определённых физиологических условиях [Wang, Gerstein, Snyder, 2009]. Этого определения будем придерживаться в дальнейшем в данной работе. Транскриптом включает в себя набор всех белок-кодирующих РНК, представленных в биологическом образце, рибосомные РНК, малые ядрышковые РНК, микроРНК, длинные некодирующие РНК и другие виды РНК [McGettigan, 2013]. Транскриптом выполняет роль связующего звена между геномом организма и его физическими характеристиками [Velculescu и др., 1997], и показывает динамику экспрессии генов [Dong, Chen, 2013], поэтому изучение транскриптома может пролить свет на особенности работы генома

Исследования транскриптома начались с появлением технологии ПЦР в реальном времени, позволявшей количественно оценить уровень представленности конкретного транскрипта в биологическом образце. Эта технология является отличается высокой точностью, но является времязатратной, так как позволяет оценивать экспрессию только отдельных генов. Кроме того, для проведения ПЦР в реальном времени необходимо иметь информацию о последовательности исследуемого гена. Другой экспериментальной технологией оценки уровня транскрипции генов является серийный анализ экспрессии генов (SAGE, Serial Analysis of Gene Expression) [Velculescu и др., 1995]. Затем, на протяжении долгого времени доминирующей технологией в транскриптомных исследованиях были микрочипы [Schena и др., 1995]. Микрочипы оценивают экспрессию одновременно десятков тысяч генов, что позволяет исследовать весь транскриптом организма одновременно. Однако, чтобы исследовать экспрессию гена с помощью микрочипа, необходимо знать последовательность этого гена в момент разработки чипа. Кроме того, микрочипы не позволяют определять последовательности исследуемых генов и обнаруживать новые полиморфные локусы.

В 2005 году была анонсирована первая платформа для высокопроизводительного секвенирования (High-throughput sequencing или next-generation sequencing, NGS) – 454 GS

FLX Genome Analyzer [Margulies и др., 2006]. В 2006 году была опубликована первая статья [Cheung и др., 2006], описывающая применение NGS для секвенирования транскриптома. Технология, основанная на выделении и секвенировании тотальной мРНК образца, была названа RNA-seq. RNA-seq выгодно отличается от микрочипов тем, что не требует предварительного знания изучаемых последовательностей, а значит, может быть использована для изучения транскриптомики немодельных организмов. Чувствительность RNA-seq выше, чем чувствительность микрочипов. RNA-seq даёт более точные количественные данные об уровнях экспрессии генов и изоформ. Наконец, RNA-seq требует меньшего количества РНК для проведения эксперимента [Mantione и др., 2014]. Как следствие, эта технология заняла место микрочипов как наиболее эффективного метода транскриптомных исследований [Shendure, 2008]. RNA-seq используется для поиска дифференциальной экспрессии генов и изоформ, поиска новых генов, обнаружения полиморфизмов, секвенирования транскриптомов немодельных организмов.

На сегодняшний день существует большое разнообразие платформ для секвенирования нового поколения, методов подготовки материала для секвенирования и конкретных приложений метода. Вариации RNA-seq включают секвенирование микро-РНК [Aldridge, Hadfield, 2012], профилирование рибосом [Ingolia и др., 2009], секвенирование мРНК с 3' конца и многие другие. Отметим, что всё большее развитие получает секвенирование третьего поколения с использованием платформ Pacific Biosciences и Nanopore, что делает их актуальными для точной идентификации вариантов сплайсинга [Bleidorn, 2016; Gupta и др., 2015]. Однако, данный обзор сфокусирован на классической технологии RNA-seq, биоинформатическом анализе данных секвенирования и применении технологии к исследованиям растительных транскриптомов.

1.6.1 Платформы для секвенирования второго поколения

Платформы секвенирования второго поколения получили наиболее широкое распространение в последние годы, кроме того, следует отметить, что они использовались для генерации экспериментальных данных в настоящей работе, поэтому в обзоре мы уделим им основное внимание. На данный момент существует несколько производителей платформ для высокопроизводительного секвенирования. Наиболее распространены платформы таких производителей, как Illumina, Applied Biosystems, Roche, BGI, Qiagen,

Thermo Fisher Scientific. Они различаются по принципу действия и имеют свои особенности. Рассмотрим некоторые из этих платформ подробнее.

Технология секвенирования Illumina

Платформы Illumina традиционно используют секвенирование парных последовательностей. Для этого секвенируемый геном/транскриптом фрагментируется, и к концам полученных фрагментов пришиваются разные адаптеры. Далее последовательности иммобилизуются на подложке, на которой закреплены олигонуклеотиды, комплементарные каждому из адаптеров. Последовательности амплифицируются в ходе мостиковой ПЦР до образования кластеров идентичных последовательностей, где вперемешку тесно расположены прямые и комплементарные последовательности. Соответственно, в зависимости от направления цепи каждой конкретной молекулы ДНК, один из двух адаптеров будет комплементарно связан с закреплёнными на подложке олигонуклеотидами, в то время как второй будет находиться на свободном конце молекулы. Далее в камеру секвенирования добавляются праймеры, комплементарные одному из адаптеров. Они связываются с адаптерами, расположенными на свободных концах, и происходит секвенирование в ходе синтеза. После прохождения реакции синтезированные комплементарные цепи удаляются из реакционной смеси, и добавляются праймеры, комплементарные второму адаптеру. Реакция секвенирования повторяется. Таким образом, для каждого из кластеров последовательностей, содержащих прямые и комплементарные цепи ДНК, записываются одна за другой две последовательности, соответствующие концам длинного фрагмента. Исходя из известных данных о методе фрагментации, можно судить о том, на каком удалении друг от друга в геномной последовательности или мРНК находятся секвенированные фрагменты. Большинство платформ позволяют секвенировать также и непарные прочтения, в этом случае просто пропускается стадия секвенирования последовательности со второго адаптера. Это ускоряет процесс секвенирования, но в результате секвенируются только короткие фрагменты РНК, что затрудняет сборку транскриптома *de novo* на их основе. Процесс секвенирования нуклеиновых кислот на платформе Illumina схематично показан на рисунке 3.

При секвенировании на платформе Illumina может возникать ошибка из-за того, что сигнал, испускаемый одним кластером молекул, возбуждает два детектора одновременно.

Это приводит к дупликации отдельных прочтений, то есть, фактически, некоторые последовательности нуклеиновых кислот оказывают секвенированы два раза вместо одного. Такой эффект называют оптической дупликацией [Klepikova и др., 2017]. Чтобы решить эту проблему, используют квазислучайные олигонуклеотидные метки, пришиваемые к секвенируемым последовательностям [Arnaud и др., 2016]. В результате, обнаружив одинаковые прочтения с одинаковыми метками, можно с уверенностью утверждать, что это – оптические дупликации, и удалять их из библиотек. Также, с дупликациями можно бороться биоинформатическими методами.

Первой из платформ линейки, разработанной компанией Solexa, стала платформа Genome analyzer. Эта платформа секвенировала последовательности длиной по 30-35 оснований на каждой из концов фрагмента ДНК. Пробоподготовка занимала около 6 часов, после чего работа секвенатора требовала от 2 до 3 суток. Платформа генерировала до 1 миллиарда пар оснований за один запуск. Из-за малой длины фрагментов прочтений платформа использовалась для ресеквенирования [Fox, Filichkin, Mockler, 2009], поиска полиморфизмов [Trick и др., 2009], создания транскриптомных профилей [Wu и др., 2010], поиска микроРНК [Chen и др., 2009b], ChIP-seq [Johnson, Mortazavi, Myers, 2007; Lefrançois, Masopust, 2009], однако были и опыты секвенирования и сборки бактериальных геномов [Farrer и др., 2009].

В дальнейшем компания Solexa была приобретена компанией Illumina, после чего секвенаторы этой линейки выпускались именно под брендом Illumina. Наиболее новая из разработанных компанией Illumina платформ – NovaSeqX. Эта платформа генерирует до 6 триллионов пар оснований и 20 миллиардов парных прочтений длинами до 250 оснований за один запуск, требуя на это менее двух суток [emea.illumina.com].

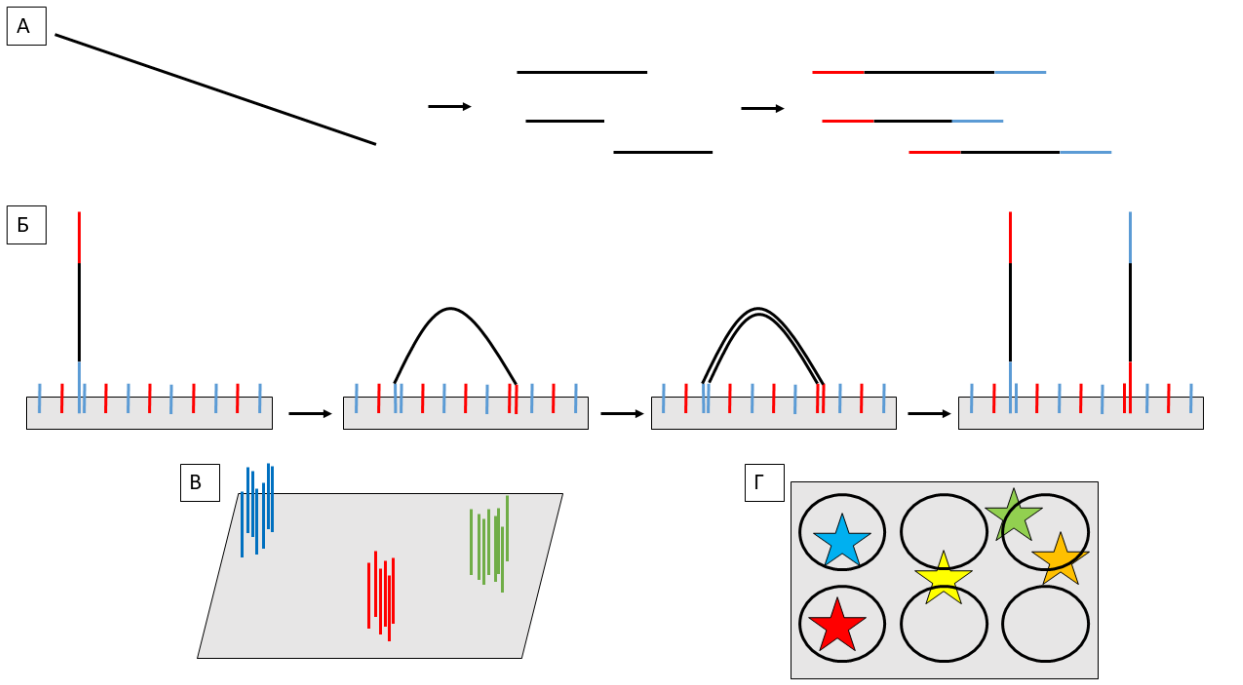

Рис. 3. Схема секвенирования нуклеиновых кислот на платформе Illumina. (А) – подготовка биологического материала, фрагментация кДНК и пришивание адаптеров. (Б) – иммобилизация фрагментов на подложке, амплификация с помощью ПЦР по принципу «мостика». (В) – подложка с кластерами амплифицированных фрагментов. (Г) – схематичное представление регистрации детекторами световых сигналов от кластеров, слева направо, сверху вниз: два детектора, воспринимающих чёткие цветовые сигналы; два детектора, воспринимающих сигналы от одного и того же кластера, за счёт чего создаются оптические дупликации; детектор, воспринимающий сигналы от двух кластеров одновременно, из-за чего создаются помехи; детектор, не принимающий сигналов при амплификации.

Технология секвенирования Roche 454

Платформа Genome Sequencer, созданная в 2005 году компанией 454, позднее приобретённой компанией Roche, опирается на принципы пиросеквенирования [Langaee, Ronaghi, 2005]. Первоначально секвенируемая ДНК или кДНК разделяется на фрагменты длинами до 1 тысячи пар оснований, после чего к фрагментам пришиваются адаптеры, и они иммобилизуются на микрогранулах. Микрогранулы затем вносятся в отсеки объёмом около 75 пиколитров, каждый из которых может содержать только одну микрогранулу. В

ходе эмульсионного ПЦР происходит амплификация каждого из таких иммобилизованных фрагментов. После этого на плашку, на которой находятся отсеки с микрогранулами, поочерёдно наносятся необходимые для секвенирования компоненты. Нуклеотиды добавляются в реакционную смесь по отдельности. При пришивании очередного нуклеотида к секвенируемому фрагменту выделяется пирофосфат и излучается световой сигнал, который улавливается сенсором. Такой сигнал свидетельствует о присоединении нуклеотида к синтезируемой цепи. Так как нуклеотиды добавляются по одному, система может определить, пришивание какого именно нуклеотида произошло в каждом из отсеков.

В первом эксперименте за один запуск, длившийся около 4 часов, было секвенировано около 300 тысяч прочтений средней длиной около 110 оснований, что суммарно составило 33,5 миллиона прочитанных оснований [Margulies и др., 2006].

По сравнению с платформами SOLiD и Illumina технологии 454 позволяли получать прочтения большей длины, поэтому они позиционировались как наиболее подходящие для *de novo* сборки геномов и транскриптомов.

Технология IonTorrent

Платформа IonTorrent Proton/PGM sequencer отличается от остальных платформ NGS по принципу своей работы. Секвенаторы IonTorrent улавливают не световой сигнал, а изменение pH в ячейке, содержащей микрогранулу с пришитыми копиями секвенируемого фрагмента генома/транскриптома. Принципиально подготовка проб к секвенированию схожа с таковой для платформы 454 – ДНК или РНК фрагментируется, к ней пришиваются адаптеры, при помощи которых затем фрагменты иммобилизуются на микрогранулах. Гранулы разносятся по микроскопическим ячейкам на подложке. Происходит амплификация фрагментов с помощью ПЦР, в результате чего каждая микрогранула несёт на себе около 1 млн. копий одного фрагмента. На плашку с ячейками, в которых находятся микрогранулы, поочерёдно вносятся растворы нуклеотидов каждого из четырёх типов, встречающихся в последовательностях ДНК – аденина, тимина, цитозина и гуанина. При встраивании в синтезируемые комплементарные цепи нуклеотида выделяется ион H^+, и локальное изменение pH улавливается системой, из чего делается вывод о том, какой именно нуклеотид был присоединён к каждому конкретному фрагменту.

В результате секвенирования по технологии IonTorrent Proton/PGM длина секвенируемых фрагментов больше, чем по сравнению с такими платформами, как Illumina (до 600 нуклеотидов, средняя длина около 180). Однако эти прочтения имеют более высокую частоту ошибок при определении нуклеотидов. Кроме того, размер секвенируемых библиотек коротких прочтений, создаваемых платформой, сравнительно невелик (миллионы одиночных прочтений по сравнению с десятками или сотнями миллионов парных прочтений, создаваемых платформами Illumina). Это делает платформу IonTorrent подходящим средством для секвенирования фрагментов для сборки транскриптомов/фрагментов генома и поиска ДЭГов и новых генов/изоформ, но она оказывается менее подходящей для поиска SNP. Кроме того, платформа страдает от более низкого качества определения слабо представленных транскриптов.

Технология ABi SOLiD

Отдельно следует рассмотреть платформу SOLiD, разработанную компанией Applied Biosystems, позднее приобретённой компанией Invitrogen. Отличительной особенностью платформы является биохимия секвенирования, реализуемая при помощи лигазной реакции [Cloonan и др., 2008]. Вместо определения каждого отдельного нуклеотида происходит определение пары идущих последовательно нуклеотидов. В секвенаторе используются светящиеся метки четырёх цветов. Каждому из цветов соответствуют четыре пары нуклеотидов, различающиеся как первой, так и второй позицией. В качестве стартовой точки секвенирования всегда используется нуклеотид тимин, входящий в состав адаптера, пришиваемого к олигонуклеотидам при подготовке к секвенированию. Таким образом, при известном первом нуклеотиде (тимин, принадлежащий к адаптеру) полученную последовательность цифр возможно перевести в последовательность нуклеотидов, однако обычно такой операции не проводится, так как каждая ошибка секвенирования в одной из позиций в середине фрагмента приведёт к тому, что оставшаяся часть фрагмента будет перекодирована ошибочно. Вместо этого, для выравнивания фрагментов на матрицу референсного генома создаются индексы генома, представляемые в том же формате последовательности цифр, каждая из которых обозначает идущие подряд два нуклеотида.

Достоинством этой платформы является высокая точность секвенирования, обусловленная особенностями секвенирования – каждый нуклеотид фактически прочитывается дважды. Это также позволяет легко отличить ошибки секвенирования от настоящих однонуклеотидных полиморфизмов, так как в первом случае от референсной последовательности отличается только один из двух сигналов, содержащих данный нуклеотид, во втором же – они оба. Из-за этого SOLiD во много позиционировалась именно как платформа для поиска ОНП. Однако именно эта специфичность платформы SOLiD является и одним из её недостатков. Действительно, обработка данных RNA-seq требует больших вычислительных мощностей и эффективных алгоритмов. При этом платформа SOLiD – единственная, записывающая секвенируемую последовательность в цветовом коде, и перекодирование в стандартные нуклеотидные является нежелательным шагом из-за описанной выше возможности получения полностью ошибочной последовательности. Как нетрудно понять, программное обеспечение, обрабатывающее стандартные данные секвенирования, напрямую с данными SOLiD работать не может. Не все разработчики реализуют в своих продуктах опцию обработки данных в формате цветового кода. Из-за этого программное обеспечение, обрабатывающее данные этой платформы, менее развито по сравнению с программным обеспечением, способным обрабатывать данные других платформ в формате последовательности нуклеотидов.

Технология BGI

BGI group была основана в 1999 году для участия в проекте секвенирования генома человека. В 2015 году компания выпустила секвенатор BGISEQ-100, после чего последовали платформы BGISEQ-200 и BGISEQ-500. Также, MGI - подразделение компании BGI – создали альтернативную ветку секвенаторов MGISEQ-2000 и MGISEQ-200 и MGISEQ-T7. Секвенирование на платформах BGI осуществляется с помощью технологии DNA nanoball sequencing, которая заключается в лигировании адаптеров к обоим концам фрагмента ДНК или РНК. Платформы поддерживают секвенирование как одноконцевых, так и парных фрагментов различных длин. Продолжительность одного запуска составляет до 48 часов.

Технология Qiagen

В 2011 году компания Intelligent Biosystems выпустила платформу для секвенирования MAX-seq. В 2012 году Intelligent Biosystems были приобретены компанией Qiagen, которая затем выпустила свою платформу GeneReader, основанную на платформе MAX-seq. Действие этой платформы основано на секвенировании путём синтеза. Фрагменты ДНК или РНК иммобилизуются на микроскопических гранулах, после чего через реакционную смесь промываются растворы нуклеотидов, модифицированных флюорофорами, специфичными для каждого из четырёх нуклеотидов. После встраивания нуклеотида в синтезируемую цепь раствор с остальными нуклеотидами вымывается, и происходит отщепление флюорофора, что сопровождается световым сигналом, улавливаемым детектором.

По состоянию на 2019 год следует отметить, что платформы 454 и SOLiD используются крайне редко. Платформы Ion Torrent, GeneReader и BGI также занимают сравнительно малую часть рынка. Таким образом, Illumina доминирует на рынке платформ для NGS. С одной стороны, это обеспечивает некоторую стандартизацию для большинства производимых исследований, и унифицирует требования к программному обеспечению. Однако есть мнение, что другие платформы, в частности IonTorrent, могут лучше, чем Illumina, подходить для решения некоторых специфических задач в отдельных RNA-seq экспериментах [Lahens и др., 2017]. Кроме того, было показано, что, при секвенировании модельного растения *A. thaliana*, по качеству производимых результатов BGISEQ-500 не уступает Illumina HiSeq4000, и, следовательно, может использоваться для решения тех же задач не с меньшей эффективностью [Zhu и др., 2018].

1.6.2 Ход эксперимента RNA-seq

Поскольку обработка и анализ экспериментальных данных RNA-seq требуют больших временных затрат, а само проведение эксперимента - материальных вложений, большой важностью обладает стадия планирования эксперимента [Conesa и др., 2016]. Сюда входит выбор платформы секвенирования, метода пробоподготовки, количества секвенируемых образцов, включая количество биологических повторностей, и, конечно, выбор собственно материала для секвенирования, то есть биологического объекта, стадии

жизненного цикла, метода обработки для получения желаемого результата, выбор органа или ткани для выделения РНК для последующего секвенирования.

В первых экспериментах RNA-seq биологические повторности могли отсутствовать, в некоторых случаях применялись технические повторности. Однако с течением времени стала очевидной необходимость использования биологических повторностей для создания выборки значений экспрессии каждого гена в каждом биологическом образце.

В литературе встречается точка зрения о предпочтительности большего количества биологических повторностей и большей длины прочтений перед глубиной секвенирования, то есть размером библиотек [Robles и др., 2012; Wang и др., 2011a]. При этом другие авторы указывают, что, хотя малой глубины секвенирования и достаточно для подсчёта высоко представленных транскриптов, увеличение глубины секвенирования позволяет точнее подсчитывать уровни транскрипции генов со слабой экспрессией [Hou и др., 2013]. Таким образом, поскольку существующие платформы, в первую очередь Illumina, предоставляют с каждой следующей машиной для секвенирования увеличение количества секвенируемых прочтений, и вкупе с высокой стоимостью каждого запуска секвенатора, в настоящее время в большинстве экспериментов используются три повторности каждого биологического образца.

Мультиплексинг секвенируемых библиотек может снизить стоимость получения одной библиотеки [Smith и др., 2010]. При мультиплексинге в процессе пробоподготовки к фрагментам кДНК пришиваются олигонуклеотиды, уникальные для каждой отдельной библиотеки – "штрих-коды". После секвенирования каждое прочтение может быть отнесён к конкретной библиотеке на основании своего штрих-кода. Это позволяет секвенировать большое количество библиотек за один запуск секвенатора. Для наибольшей точности результатов нужен баланс между глубиной секвенирования и количеством биологических повторностей, детали которого зависят от конкретного эксперимента, однако в любом случае лучше брать 4 или больше повторностей [Lamarre и др., 2018]. Для сравнения экспрессии генов в близкородственных организмах достаточно трёх повторностей на каждый образец, в то время как для менее схожих организмов (сравнение культурных сортов и диких сородичей, изучение растений из удалённых популяций) требуется больше биологических повторностей [Williams и др., 2014].

Далее перед исследователем встаёт проблема обогащения мРНК в изучаемых образцах. Рибосомная РНК в норме может составлять до 90% всей РНК клетки. Соответственно, для изучения экспрессии мРНК, важно удалить рРНК из смеси на стадии пробоподготовки. Зачастую, это решается поли-А обогащением при выделении РНК. Другой метод удаления рРНК из секвенируемых библиотек – использование зонд-направленной деградации (Probe-assisted degradation) [Archer, Shirokikh, Preiss, 2015]. В смесь вводятся олигонуклеотиды, комплементарные рибосомной РНК, после чего двуцепочечные фрагменты деградируются нуклеазами. В отдельных случаях помимо последовательностей, комплементарных рРНК, вводятся также последовательности, комплементарные наиболее интенсивно экспрессирующимся генам. Удаление их из секвенируемых библиотек увеличивает точность обнаружения и определения слабо представленных транскриптов.

1.6.3 Биоинформатическая обработка данных RNA-seq

Теперь, в свою очередь, рассмотрим подробно стадии биоинформатического анализа RNA-seq, алгоритмы и программные продукты, используемые на каждой из этих стадий. Также в этом разделе мы постараемся показать, что для каждой из стадий биоинформатического анализа экспериментов RNA-seq существует множество программных решений, комбинации которых дают огромное количество возможных конвейеров для обработки RNA-seq, и что производительность отдельных конвейеров зависит от входных данных. По утверждению ряда авторов, из этого следует, что в каждом конкретном случае для анализа RNA-seq необходимо подбирать конкретный конвейер, который окажется оптимальным для имеющихся данных.

1.6.3.1 Фильтрация библиотек

Прежде всего, следует обратить внимание на последовательности адаптеров. В большинстве случаев эти последовательности не секвенируются, либо они удаляются программным обеспечением в процессе разбиения прочтений по библиотекам. Тем не менее, иногда фрагменты адаптеров остаются в составе прочтений, входящих в библиотеки. В таком случае используются программные средства для удаления последовательностей адаптеров из прочтений библиотек [Payá-Milans и др., 2018]. Для этой цели служат такие

программы, как Cutadapt [Martin, 2011], Trimmomatic [Bolger, Lohse, Usadel, 2014], Fastp [Chen и др., 2018]. Отметим, что эти программы совмещают в себе функции удаления адаптеров и фильтрации библиотек по качеству, длинам и прочим параметрам (см. далее).

В процессе секвенирования каждому прочитанному нуклеотиду (или, в случае платформы SOLiD, каждой паре последовательных нуклеотидов) присваивается значение качества Phred quality score [Ewing, Green, 1998], рассчитываемое по следующей формуле:

$$Phred = -\log_{10} P_e \quad (1)$$

где P_e – вероятность ошибочного прочтения нуклеотида. Таким образом, значение качества 20 означает, что нуклеотид был прочитан верно с вероятностью 0,99; значение 30 – что нуклетид был прочитан правильно с вероятностью 0,999, и так далее.

Часто исследователи удовлетворяются удалением из библиотек слишком коротких и низкокачественных прочтений и прочтений, содержащих слишком большое количество неопределённых нуклеотидов [Liu и др., 2016; Tombuloglu и др., 2015; Yang и др., 2015]. Пороги длин, качества и содержания неопределённых нуклеотидов могут варьировать в зависимости от конкретных данных и целей эксперимента, но чаще всего удаляются прочтения со средним качеством ниже 20, длиной ниже 50 нуклеотидов и содержанием неопределённых нуклеотидов выше 10%. В случае секвенирования библиотек парных прочтений удаляются также прочтения, не имеющие пары. Однако в некоторых случаях проводится более тщательный контроль качества библиотек: коррекция k-меров, удаление дупликаций, удаление рибосомной РНК и других некодирующих видов РНК, удаление контаминантов в виде генетического материала человека, *Escherichia coli* и других организмов.

При выборе стратегии фильтрации следует оценивать различные параметры, так как фильтрация только по одному какому-либо параметру может повлиять на качество дальнейшего анализа не лучшим образом [Williams и др., 2016]. Так, если удалять по 10 первых оснований каждого из прочтений, не принимая во внимание значения качества, то количество обнаруженных ДЭГ снижается на 2% [Amaral и др., 2014]. Имеет смысл удалять прочтения со слишком маленькими длинами (меньше 20 нуклеотидов), что особенно верно при работе с платформами IonTorrent, которые, в отличие от платформ Illumina, создают прочтения с некоторым разбросом длин [Williams и др., 2016]. Другая особенность платформ IonTorrent – постепенное падение качества по мере приближения к 5'-концу

прочтения, что тем более заметно, чем выше длины прочтений. Так, для прочтений с длинами более 300 нуклеотидов доля нуклеотидов с качеством ниже 20 может стать весьма заметной. Однако, эта проблема может быть легко разрешима, так как большинство программ для фильтрации качества предоставляют возможность удалять нуклеотиды с 5’ конца прочтения после позиции последовательности с определенным порядковым номером, выбранным пользователем, или же нуклеотиды, имеющие качество Phred ниже заданного.

В целом, хотя фильтрация понижает общее количество прочтений в библиотеках, после фильтрации возрастает доля успешно картированных прочтений [Del Fabbro и др., 2013].

1.6.3.2 Картирование прочтений на референсный геном

К моменту появления технологии RNA-seq было разработано большое количество программ, служащих для выравнивания последовательностей между собой – Blast [Altschul и др., 1990], BLAT [Kent, 2002], MAQ [Li, Ruan, Durbin, 2008a]. Однако, когда появилась задача выравнивания миллионов последовательностей с длинами порядка нескольких десятков нуклеотидов на последовательность ДНК размером в полный геном (т.е. порядка сотен миллионов или даже миллиардов нуклеотидов), стало очевидно, что имеющиеся алгоритмы не подходят ни по показателям качества выравнивания, ни по производительности и быстродействию.

Картированием называется выравнивание каждого прочтения библиотеки на референсную (полногеномную) нуклеотидную последовательность. Поскольку каждое прочтение получается в результате секвенирования фрагмента какой-либо конкретной мРНК, который, в свою очередь, получается в результате транскрипции какого-либо конкретного гена, то для каждого прочтения теоретически можно однозначно определить положение в геноме, которому он соответствует. В действительности это осложняется наличием повторов, гомологичных генов, ошибками секвенирования. Схематично картирование библиотек RNA-seq представлено на рисунке 4.

Большинство мРНК в процессе созревания проходят стадию сплайсинга, в ходе которой из транскрипта удаляются интроны. Интроны могут иметь длины до десятков или даже сотен тысяч нуклеотидов. В результате, позиции, стоящие по соседству в

последовательности мРНК, могут быть разнесены на большие расстояния в последовательности генома. Это значит, что, если прочтение попадает на стык двух экзонов в мРНК, для его успешного картирования на геном необходимо определить два участка генома по краям интрона, на которые будут картированы два подряд идущих фрагмента прочтения. Первые платформы для NGS, однако, имели низкие длины прочтений в библиотеках (25 нуклеотидов для SOLiD и Illumina/Solexa). Это отражалось, в частности, в малой доле прочтений, попадавших на стыки экзонов. Поэтому ранние программы для картирования – MAQ [Li, Ruan, Durbin, 2008b], Bfast [Homer, Merriman, Nelson, 2009], BWA [Li, Durbin, 2009], Bowtie [Langmead и др., 2009] и Bowtei2 [Langmead, Salzberg, 2012], Shrimp [Rumble и др., 2009] не учитывали требование к выравниванию прочтений с учётом сплайсинга. Однако с ростом длин прочтений увеличивается доля прочтений, попадающих на стык экзонов, вплоть до того, что картирование без учёта сплайсинга дает бессмысленные результаты [Lindner, Friedel, 2012]. Были разработаны программы – STAR [Dobin и др., 2013], TopHat [Trapnell, Pachter, Salzberg, 2009] и TopHat2 [Kim и др., 2013a], Hisat и Hisat2 [Kim, Langmead, Salzberg, 2015], GSNAP из пакета программ GMAP [Wu, Watanabe, 2005], SeqSaw [Wang и др., 2011a], названные обобщённо “сплайсинг-совместимые” (splice-aware aligners), которые способны картировать прочтения на геном по частям, обнаруживая случаи сплайсинга.

Вкратце, для картирования прочтений используются так называемые начальные выравнивания (‘seed’), которые должные точно совпадать в прочтении и референсе. После обнаружения для каждого прочтения совпадения с каким-либо начальным выравниванием, программа пытается удлинить выравнивание поиском совпадения соседних нуклеотидов в прочтении и референсе. Для этого служат хэшированные таблицы или деревья суффиксов [Li, Homer, 2010]. В целом, считается, что алгоритмы, основанные на деревьях суффиксов, превосходят алгоритмы с хэшированными таблицами по быстродействию, но уступают в точности [Mielczarek, Szyda, 2016].

Burrows-Wheeler Aligner (BWA) и BowTie используют преобразование Барроуза-Уилера для сравнения строк. STAR использует так называемый максимальный картируемый префикс, являющийся адаптацией максимального точного совпадения (Maximum Exact Match, MEM), используемого во многих других программах, например, Mummer [Kurtz и др., 2004] и MAUVE [Darling и др., 2004]. Максимально точное

совпадение – гомологичный участок между прочтением и референсом, который невозможно продлить без появления несоответствий.

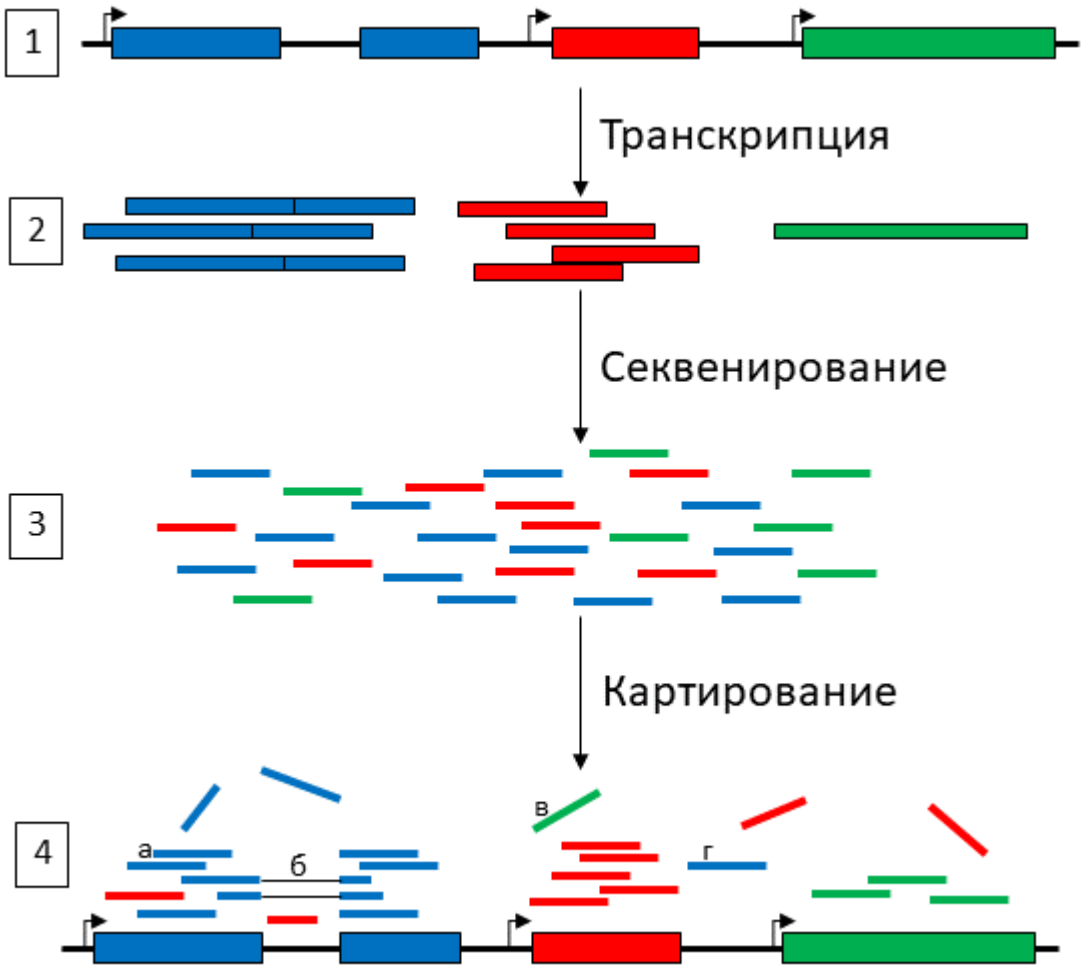

Рис. 4. Картирование библиотек коротких прочтений на референсный геном. Цифрами обозначены: 1 – фрагмент генома организма, содержащий три гена; 2 – матричная РНК, транскрибированная с трёх представленных генов; 3 – библиотеки коротких прочтений, полученные с помощью процедуры RNA-seq; 4 – картирование коротких прочтений на референсный геном организма. Буквами обозначены прочтения: а – картированные правильно без учёта сплайсинга; б – картированные правильно с учётом сплайсинга; в – не картированные; г – картированные ошибочно.

TopHat [Kim и др., 2013a] использует двухстадийный алгоритм – сначала при помощи Bowtie [Langmead и др., 2009] или Bowtie2 [Langmead, Salzberg, 2012] картируются все прочтения, которые возможно картировать без сплайсинга. Затем для оставшихся прочтений проводится картирование с учётом сплайсинга. Этот алгоритм считается достаточно быстрым, но менее точным, чем аналоги. Схожей стратегии придерживаются RUM [Grant и др., 2011] и RNASEQR [Chen и др., 2012b].

Dart [Lin, Hsu, 2018b] – программа для картирования, использующая алгоритм разбивки, ранее применявшийся в программе Kart [Lin, Hsu, 2017] для картирования последовательностей, разработанной теми же авторами. Dart оперирует локальными максимальными точными совпадениями (Local MEM, LMEM). Программа ищет максимально длинный фрагмент с начала прочтения, который можно полностью выровнять на референс без вставок и замен. Затем, операция повторяется, начиная со следующего нуклеотида прочтения. Все обнаруженные так совпадения сортируются по их положению в геноме. Если два соседних совпадения разделены расстоянием меньше, чем означенная пользователем или заданная по умолчанию максимальная длина интрона, то программа предполагает, что эти совпадения отражают истинное положение прочтения в референсе.

Отдельные исследования были посвящены сравнению производительности программ для картирования [Engström и др., 2013; Benjamin и др., 2014; Williams и др., 2017]. STAR обычно высоко оценивают в плане быстродействия и точности [Engström и др., 2013; Williams и др., 2017]. GSNAP также выделяют как программу, имеющую высокую точность, хотя и низкую скорость по сравнению со STAR [Engström и др., 2013; Payá-Milans и др., 2018]. При этом, однако, утверждается, что, хотя влияние программ для картирования может быть очень велико, следует рассматривать их не отдельно, а в свете остальных программ, вовлечённых в анализ данных RNA-seq [Williams и др., 2017].

В результате картирования прочтений на геном большинство программ формируют файлы в формате SAM (Sequence Alignment/mapping). Это текстовый табличный формат, где каждая строка соответствует картированию одного прочтения на одну позицию в геноме, и отражает хромосому, локализацию на хромосоме того места, куда было картировано прочтение, качество картирования и некоторые другие параметры.

Полученное картирование библиотек может быть проанализировано по ряду показателей для оценки качества.

1.6.3.3 *de novo* реконструкция транскриптома

Для анализа транскриптома организма путём картирования библиотек RNA-seq на референсный геном, как описано выше, необходимо, чтобы последовательность генома организма была секвенирована, должным образом реконструирована и доступна для пользователей. Несмотря на то, что количество секвенированных геномов постоянно растёт, этот метод остаётся недоступным для огромного количества не-модельных организмов, чьи геномы ещё не были секвенированы [Fu и др., 2018]. Для исследования транскриптомов таких организмов используют такой метод, как *de novo* сборка транскриптома [Wang, Gribskov, 2017]. Суть этой процедуры состоит в реконструкции полноразмерных последовательностей транскриптов из прочтений библиотек. Кроме того, *de novo* сборка позволяет даже для хорошо изученных модельных организмов обнаруживать новые транскрипты, которые по каким-либо причинам не попали в последовательности раннее секвенированного транскриптома или генома [Honaas и др., 2016].

В настоящее время существуют два подхода к реконструкции последовательностей транскриптома из библиотек коротких прочтений. Первый – метод перекрывающихся последовательностей (OLC, Ovelap-Layout-Consensus), второй – метод графов де Брёйна [Li и др., 2012; Schliesky и др., 2012]. Метод перекрывающихся последовательностей заключается в попарном выравнивании прочтений и создании ориентированных графов, где каждый узел – это одно прочтение. В качестве рёбер выступают перекрывания между прочтениями. Таким образом, путь по графу показывает последовательность, которую можно составить из перекрывающихся прочтений. Метод перекрывающихся графов больше подходит для сборки контигов из сравнительно малого количества прочтений большой длины с большими участками перекрывания, и, таким образом, используется чаще для сбора последовательностей, полученных методом Сэнгера, или методами секвенирования третьего поколения [Cui и др., 2020].

Второй метод заключается в создании графа де Брёйна, в котором вершинами выступают k-меры, то есть последовательности нуклеотидов заданной длины k. Затем на графе отмечаются все пути, дающие в результате имеющиеся в используемых библиотеках короткие прочтения. После этого отмечают все пути, содержащие непрерывные последовательности перекрывающихся прочтений. Таким образом, находят

последовательности контигов, которые можно собрать из прочтений библиотеки. Этот метод используется во многих программах-сборщиках транскриптома, как, например, Trinity [Grabherr и др., 2013], Trans-ABySS [Robertson и др., 2010], SOAP-denovo Trans [Xie и др., 2014], Velvet-Oases [Schulz и др., 2012].

Для сборщиков, основанных на методе графов де Брёйна, существует важный параметр k – длина k-мера, который сборщик использует для создания графа де Брёйна. Этот параметр может устанавливаться пользователем при запуске. Увеличение k повышает точность сборки, но одновременно увеличивает сложность вычисления [Fu и др., 2018]. Однако, при более высоких k сборщик может не обнаружить пересечение между прочтениями, если длина этого пересечения меньше k. Зачастую используется следующая стратегия – проведение предварительных сборок при разных значениях k, после чего из полученных сборок путём конкатенации и удаления избыточности (см. ниже) составляется «мета-сборка» [Wang, Gribskov, 2017]. Пример графов де Брёйна проиллюстрирован на рисунке 5.

Поскольку на сегодняшний день разработано достаточно большое количество программ, осуществляющих сборку транскриптома *de novo*, отдельные исследования были посвящены вопросу о производительности и точности этих сборщиков. Обзоры, рассматривающие сравнение нескольких программ для сборки транскриптома de novo, как правило, выделяют в качестве лучших из них и наиболее популярных Trinity [Grabherr и др., 2013], SOAPdenovo-trans [Xie и др., 2014], Velvet-Oases [Schulz и др., 2012]. Trinity, помимо непосредственно сборщика, включает в себя также большой набор утилит для оценки качества сборки, удаления слабо представленных контигов, и других манипуляций с *de novo* сборкой. SOAPdenovo-trans отмечают как программу, подходящую для сборки растительных транскриптомов [Payá-Milans и др., 2018].

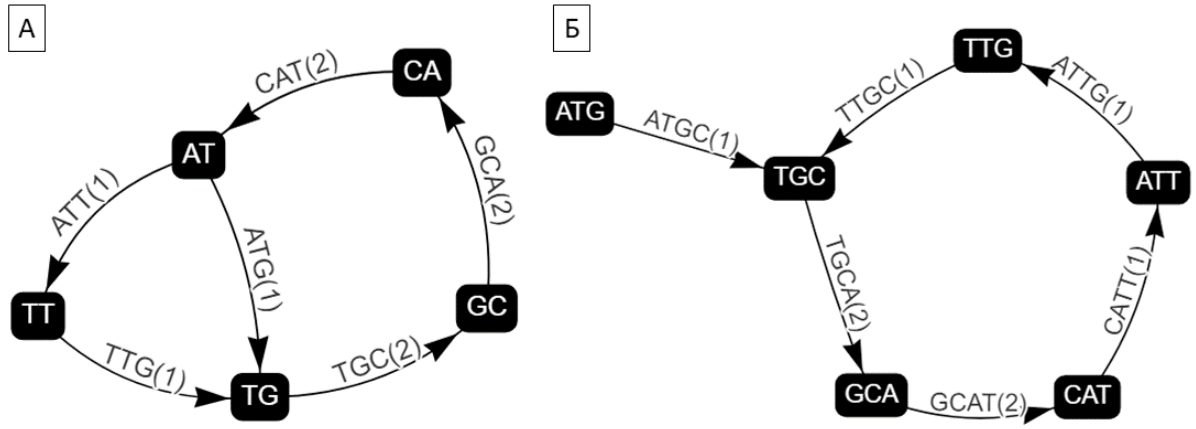

Рис. 5. Графы де Брёйна, путь по которым даёт последовательность ATGCATTGCAT. Длины k-мера в данном случае равны: (А) двум, (Б) трём. Иллюстрации подготовлены с помощью сервиса De-Bruijn-Graph-Genome-Visualization [https://github.com/schostac/De-Bruijn-Graph-Genome-Visualization].

Наконец, существует мнение, что использование нескольких сборщиков и дальнейшее создание одной «мета-сборки» дополнительно улучшает чувствительность и точность метода [Cerveau, Jackson, 2016]. Под мета-сборкой при этом понимается совокупность всех *de novo* собранных разными программами контигов после удаления избыточности (см. ниже)

1.6.3.4 Анализ качества сборки

Чтобы оценить качество сборки транскриптома, необходимо, прежде всего, разработать критерии для оценки этого качества. Традиционно использовались критерии, пришедшие из практики сборки геномов *de novo* – количество контигов, средняя длина и сопутствующие ей параметры N50 и L50, а также среднее покрытие контигов прочтениями библиотек. В геномных сборках увеличение длины контига и его покрытия говорит о более высоком качестве. Однако в случае транскриптомных сборок, само по себе увеличение количества контигов и значения N50 не отражает увеличение качества сборки транскриптома, так как реальное количество транскриптов и их длины в исследуемом транскриптоме могут быть неизвестны [O'Neil, Glowatz, Schlumpberge, 2013]. Вместо этого предлагается использовать другие параметры: доля прочтений из библиотек,

использованных для сборки, которые затем успешно могут быть картированы на эту сборку; представленость в сборке широко встречающихся генов; количество полученных в сборке унигенов [Honaas и др., 2016]. Для оценки первого параметра проводится картирование библиотек на полученную сборку какой-либо из программ, осуществляющих картирование прочтений на референсные последовательности. Специально для количественной оценки второго параметра была разработана программа BUSCO [Simão и др., 2015], которая проводит выравнивание контигов на заранее подготовленную базу данных, состоящую из последовательностей, характерных для изучаемой группы видов – животных, сосудистых растений или других. База данных при этом содержит заранее подобранные последовательности, встречающиеся у всех представителей этой группы, и доля этих последовательностей, представленных в *de novo* собранных контигах, даёт представление о полноте сборки. Количество унигенов оценивается после устранения избыточности сборки.

Избыточность сборки – наличие в ней контигов, представляющих фрагменты или разные формы одних и тех же транскриптов. При этом сборка разных изоформ гена не создаёт избыточности, в отличие от артефактов сборки, являющиеся фрагментами целых транскриптов. Избежать появления избыточных контигов в сборке практически невозможно [Yang, Smith, 2013]. Для удаления избыточных контигов существуют такие программы как Uclust из пакета Usearch [Edgar, 2010], и CD-Hit [Fu и др., 2012]. Контиги, оставшиеся после удаления избыточных, часто называют «унитиги». Количество этих контигов и наличие в них открытых рамок считывания (поскольку происходит реконструкция последовательностей транскриптов, имеющих рамки считывания) – один из показателей качества сборки *de novo*.

Фактор, создающий существенные сложности для сборки транскриптомов *de novo* – контаминация чужеродной ДНК или РНК. Контиги, собранные из чужеродного материала могут искажать размер сборки и уровни экспрессии транскриптов, что сказывается негативно на оценке дифференциальной экспрессии [Schmieder, Edwards, 2011]. Существуют отдельные методы поиска и удаления контаминации, связанные с выравниванием контигов на последовательности изучаемого организма и потенциальных контаминантов [Schmieder, Edwards, 2011], выявлением аномалий в CG-составе и покрытии последовательностей [Kumar и др., 2013] или кластеризацией последовательностей и

удалении выпадающих кластеров [Lafond-Lapalme и др., 2017]. Некоторые из этих методов подходят только для модельных видов, некоторые непригодны для сборок транскриптома и хорошо используются только для сборок генома.

Ещё одним фактором, создающим сложности в сборке транскриптома *de novo*, являются собираемые химерные контиги. Это – артефакты сборки, получаемые из-за наличия у разных генов схожих доменов или гомологичных районов. Вследствие этого у прочтений, относящихся к этим генам, могут быть перекрывающиеся участки, из-за чего сборщики воспринимают их, как части одного транскрипта. Другая причина появления химер – ошибки амплификации, создающие в библиотеках прочтения, состоящие из соединенных между собой частей других прочтений. Химеры, как и контаминанты, могут создавать помехи в адекватном анализе полученной сборки. Химеры можно опознавать, картируя контиги на геном: контиги, разные части которых отстоят друг от друга на очень большое расстояние (заведомо больше длины интрона изучаемого вида) или картируются с высокой точностью на разные хромосомы – скорее всего, химерные последовательности [Honaas и др., 2016; Wang, Gribskov, 2017]. Также, контиги, разные участки которых картируются на разные CDS организма, выделяются как химеры [Armero и др., 2017].

Несколько программ были разработаны с целью удаления химерных последовательностей из данных RNA-seq – ChimeraScan [Iyer, Chinnaiyan, Maher, 2011], ChimPipe [Rodríguez-Martín и др., 2017]. Наконец, TransRate [Smith-Unna и др., 2016], программа, созданная для контроля качества *de novo* сборок, обеспечивает в том числе опознание химерных контигов. TransRate также ставит в соответствие каждому собранному контигу определённый транскрипт из числа аннотированных для данного организма, основываясь на двунаправленных лучших совпадениях (bi-directional best hit).

1.6.3.5 Оценка уровня экспрессии транскриптов

Как было сказано выше, программы для картирования формируют на выходе файл с расширением «.sam» (Sequence Alignment/Mapping), в котором в виде текстовой таблицы указаны выравнивания всех прочтений библиотеки на конкретные позиции в геноме. Информация в файле SAM позволяет извлечь количество прочтений, выравненных на определенный участок генома и оценить т.н. покрытие генома такими выравниваниями (количество прочтений, выравненных на отдельный нуклеотид генома). На основании

информации о структурной разметке генов в геномной последовательности, которая представлена в файлах формата GFF (general feature format) это позволяет оценить среднее количество прочтений, выравненных на последовательность каждого гена или экзона. Такие оценки можно выполнить при помощи ряда программ, таких как bedtools [Quinlan, Hall, 2010], HTSeq-count [Anders, Pyl, Huber, 2015], featureCounts из пакета subread [Liao, Smyth, Shi, 2014], RSEM [Li, Dewey, 2011].

1.6.3.6 Нормализация уровней экспрессии

Множество факторов вносят искажение в наблюдаемые в экспериментах RNA-seq величины экспрессии генов. В качестве таких факторов упоминают размеры библиотек, длины генов, GC-состав генов. Для дальнейшей обработки данных и проведения анализа полученных результатов необходимо устранить воздействие этих факторов. Для этого проводится нормализация значений экспрессии, то есть приведение значений экспрессии к величинам, сравнимым между библиотеками и между генами.

Первоначально было высказано предположение, что оценки уровней экспрессии генов, полученные при помощи экспериментов по высокопроизводительному транскриптомному секвенированию не потребуют сложных методов нормализации [Wang, Gerstein, Snyder, 2009]. В действительности же, множество факторов могут искажать количественные оценки уровней экспрессии генов, и влияние нормализации может быть очень велико [Bullard и др., 2010]. Наиболее очевидные - размеры библиотек, длины генов, относительное содержание транскриптов наиболее интенсивно экспрессирующихся генов [Rapaport и др., 2013].

Простейший вариант нормированных значений уровней экспрессии – CPM (counts per million). В этом способе нормализация проводится только на размер библиотек. Значение CPM вычисляется по следующей формуле:

$$CPM_i = 10^6 \cdot \frac{X_i}{\sum_{j=1}^{N} X_j} \qquad (2)$$

Где i – ген, для которого подсчитывается уровень экспрессии; X_j – количество прочтений, картированных на ген j; N – общее количество исследуемых генов. Важно отметить, что в факторе размера библиотек в данном случае учитываются только те

прочтения, которые успешно картированы на геном или транскриптом исследуемого организма.

Другим фактором, влияющим на оценку уровня экспрессии, основанной на подсчете количества прочтений, является его длина. Действительно, поскольку на стадии фрагментации РНК транскрипты измельчаются, прежде чем начнётся амплификация и непосредственно секвенирование, транскрипт большей длины будет разбит на большее количество фрагментов, каждый из которых имеет шанс быть секвенированным. Следовательно, количество прочтений, секвенированных с транскрипта каждого конкретного гена, пропорционально длине данного транскритпа. Таким образом, уровни экспрессии следует нормировать также на длину генов по следующей формуле:

$$FPKM_i = 10^6 \cdot \frac{X_i}{L_i \cdot \sum_{j=1}^{N} X_j} \quad (3)$$

Где i – ген, для которого подсчитывается уровень экспрессии; L_i – длина гена i; X_j – количество прочтений, картированных на ген j; N – общее количество исследуемых генов. Полученные значения называют FPKM (Fragments Per Kilobase per Million reads) [Mortazavi и др., 2008]. FPKM используется для библиотек парных прочтений и показывает количество прочтений из каждой пары, картированных на ген. Также, приводится параметр RPKM (Reads Per Kilobase per Million mapped reads), который для парных библиотек отражает количество пар, картированных на ген, и, следовательно, соотносится с FPKM как 1:2. Для непарных прочтений значения RPKM и FPKM синонимичны.

Также приводят такой метод нормализации, как ERPKM (Effective RPKM), который отличается тем, что вместо длины транскриптов или генов приводится эффективная длина транскриптов. Эффективная длина транскрипта вычисляется следующим образом:

$$EL_{ai} = L_i - R_a + 1 \quad (4)$$

Где EL_{ai} – эффективная длина транскрипта i при нормализации библиотеки a, L_i – фактическая длина транскрипта i, R_a – средняя длина прочтения в библиотеке a. Далее нормализация проводится аналогично методу RPKM, за исключением того, что вместо длины транскрипта L_i используется эффективная длина EL_i. Эта поправка делается, чтобы учесть, что в зависимости от длины прочтения в библиотеке разная длина транскрипта может производить разный эффект. Нужно отметить, что такой способ нормализации, как RPKM или ERPKM, позволяет сравнивать уровни экспрессии генов в одном образце, в то

время как CPM позволяет сравнивать экспрессию каждого конкретного гена только между образцами. То есть, если пользователь хочет ранжировать гены по уровням экспрессии внутри одного образца или одной библиотеки, то необходимо сделать поправку на длины, такую как RPKM или ERPKM. Для сравнения уровней экспрессии между образцами, однако, такой поправки не требуется, так как длины одного гена во всех образцах одинаковы.

Также существует метод нормализации TPM (transcripts per million) [Li, Durbin, 2009]. Авторы предлагают измерять уровень экспрессии гена двумя метриками – доля нуклеотидов и доля транскриптов относительно всего транскриптома, приходящиеся на определённый ген или изоформу. Эти две величины связаны системой из двух уровнений:

$$\nu_i = \frac{\tau_i l_i}{\sum_j \tau_j l_j} \tag{5}$$

$$\tau_i = \frac{\nu_i}{l_i}\left(\sum_j \frac{\nu_j}{l_j}\right)^{-1} \tag{6}$$

Где l_i – длина гена i, ν_i – доля нуклеотидов, составленных геном i в транскриптоме, τ_i – доля транскриптов, составленных геном i в транскриптоме. В таком случае, метрика TPM будет выражена как $\tau_i \cdot 10^6$. Этот способ нормализации учитывает два фактора – размер библиотеки и длину каждого конкретного гена. Авторы также сравнивают методы нормализации TPM и FPKM и приходят к выводу, что для генов длиной 1 т.п.о. метрики TPM и FPKM эквивалентны. Величина экспрессии 1 FPKM примерно соответствует одному транскрипту на клетку у мышей [Mortazavi и др., 2008].

Был предложен ещё один метод, позволяющий избавиться от ошибки, создаваемой разницей в длинах генов: для подсчёта экспрессии оценивать количество прочтений, картированных не на целый ген, а на участок гена фиксированной длины, например, 250 нуклеотидов [Bullard и др., 2010]. Авторы показали, что этот метод устраняет связь между длиной прочтения и уровнем значимости дифференциальной экспрессии, которую они наблюдали ранее. Однако, сами авторы признают, что этот метод существенно уменьшает информативность результатов эксперимента RNA-seq. К примеру, при таком подходе нельзя наблюдать явления альтернативного сплайсинга и экспрессии изоформ.

Существует метод нормализации, основанный на предположении, что уровни экспрессии некоторых генов, в частности – генов домашнего хозяйства, должны оставаться

неизменными во всех секвенируемых образцах. В этом случае выбирается один какой-либо ген, охарактеризованный как ген домашнего хозяйства, о котором также есть достоверная информация, что этот ген не изменяет свою экспрессию в исследуемых условиях. Назовём этот ген нормировочным геном. После этого один из образцов выбирается как стандарт, и для каждого из остальных образцов коэффициент нормировки рассчитывается следующим образом:

$$K_a = \frac{X_s}{X_a} \quad (7)$$

Где X_s – уровень экспрессии нормировочного гена в образце, выбранном как стандарт, X_a – уровень экспрессии нормировочного гена в образце a. Далее уровни экспрессии каждого гена в образце, a умножаются на K_a.

Методы нормализации Median и Upper Quartile основаны на подсчёте квантилей [Bullard и др., 2010]. Для этого вычисляются медианное (в случае нормализации по медиане) значение уровня экспрессии или значение для верхнего квартиля (в случае нормализации по верхней квартили) для каждой из картированных библиотек. После этого вычисляется среднее значение среди всех квантилей, подсчитанных таким образом. Затем, для каждого образца вычисляется нормировочный коэффициент, равный отношению среднего значения к квантили, подсчитанной для этого образца:

$$K_j = \frac{g}{g_j} \quad (8)$$

Где K_j – это нормировочный коэффициент для библиотеки j, g_j – медиана (в случае нормализации по методу медиан) или третья квартиль (в случае нормализации по методу верхней квартили), g – среднее по всем подсчитанным таким образом квантилям. Далее все значения экспрессии этого образца умножаются на подсчитанный для него нормировочный коэффициент. Отмечено, что метод верхней квартили имеет лучшие результаты в случае, если большое количество генов в исследуемых образцах имеют низкие уровни экспрессии, в противном случае достаточно использовать метод медиан.

TMM (Trimmed Mean of M-values, Усечённое среднее М-значений) – метод нормализации, базирующийся на предположении, что в каждом исследуемом транскриптоме дифференциально экспрессирующимися является меньшая часть генов [Robinson, Oshlack, 2010]. В этом методе указывается, что, хотя истинный размер транскриптома, то есть общее количество молекул мРНК в биологическом образце,

неизвестен, можно оценить соотношения между размерами транскриптомов разных образцов. Для нормализации прежде всего подсчитываются значения изменения уровня экспрессии гена M_g и абсолютных уровней экспрессии гена A_g:

$$M_g^{ij} = log_2 \frac{\frac{Y_{gi}}{N_i}}{\frac{Y_{gj}}{N_j}} \quad (9)$$

$$A_g = log_2 \left(\sqrt{\frac{Y_{gi}}{N_i} \cdot \frac{Y_{gj}}{N_j}} \right) \text{ для } Y_{gi}, Y_{gj} \neq 0 \quad (10)$$

Где Y_{gi} – наблюдаемое из результатов картирования значение экспрессии гена g в образце i, N_i – общий размер библиотеки образца i. Оба значения затем усекаются, и подсчитывается взвешенное среднее значений M.

Отметим, что методы нормализации, основанные на подсчёте размеров библиотек, в ряде работ подвергаются критике [Rapaport и др., 2013; Lamarre и др., 2018; Soneson, Delorenzi, 2013]. Авторы утверждают, что наличие отдельных генов с высокими уровнями экспрессии может создавать искажение в размере библиотек, которое влияет на нормализацию. Для подсчёта коэффициента k_i нормализации для библиотеки i они предлагают следующий алгоритм:

$$k_i = median \frac{n_{ij}}{\sqrt[m]{\prod_{v=1}^{m} n_{iv}}} \quad (11)$$

Где n_{ij} – это уровень экспрессии гена i в образце j; m – количество образцов.

Исследования с целью оптимизировать подход к нормализации продолжаются в настоящее время. Так, в 2017 году был предложен двухстадийный метод нормализации [Li и др., 2017], состоящий из нормализации значений экспрессии каждого гена внутри образца и между образцами. Вначале уровни экспрессии каждого гена в каждом образце нормализуются при помощи способов Median или Upper Quartile, описанных выше. После этого уровни экспрессии каждого гена нужно разделить на значение медианы по экспрессии этого гена среди всех библиотек. Также, был предложен метод нормализации GeTMM [Smid и др., 2018], связанный с коррекцией значений экспрессии гена, полученных описанным выше методом TMM, на значение длины гена.

Многие пакеты, производящие поиск дифференциальной экспрессии генов - EdgeR, DESeq, DESeq2, DEXseq (см. далее) - проводят нормализацию уровней экспрессии

самостоятельно. Пользователю предоставляется выбор из нескольких доступных методов нормализации. Консенсус на тему того, какой из существующих методов нормализации оптимален для анализа уровней экспрессии, полученных из экспериментов RNA-seq, в настоящее время не достигнут, хотя отдельные авторы утверждаю, что методы нормализации TMM и Relative Log превосходят по результатам остальные методы [Dillies и др., 2013; Maza и др., 2013].

1.6.3.7 Поиск дифференциальной экспрессии генов

Самое распространённое применение RNA-seq – идентификация генов, которые существенно изменяют уровни экспрессии в экспериментах типа «опыт-контроль» [Drewe и др., 2013]. Такие гены называют экспрессирующимися дифференциально (дифференциально экспрессирующиеся гены, ДЭГ).

С точки зрения статистики, идентификация дифференциально экспрессирующихся генов основана на проверке нулевой гипотезы о том, что распределение уровней экспрессии всех генов одинаково, а все наблюдаемые различия между уровнями экспрессии генов в разных образцах объясняются статистическими флуктуациями. Для проверки этой гипотезы необходимо знать, какому распределению подчиняются значения экспрессии, наблюдаемые при помощи RNA-seq. Первым предположением было, что такие значения подчиняются распределению Пуассона [Marioni и др., 2008]. Распределение Пуассона также использовалось ранее, чтобы моделировать уровни экспрессии в микрочиповых экспериментах. Однако было обнаружено, что наблюдаемая в данных RNA-seq дисперсия выше, чем это ожидается для значений, следующих распределению Пуассона [Esnaola и др., 2013; Love, Anders, Huber, 2013]. Это явление было названо «избыточная дисперсия» (over-dispersion). В частности, оказалось, что эмпирическое распределение уровней экспрессии генов по результатам экспериментов высокопроизводительного секвенирования лучше описывается законом отрицательного биномиального распределения [Wu, Wang, Wu, 2013], которое в настоящий момент используется для моделирования уровней покрытия генов прочтениями RNA-seq наиболее часто [Huang, Niu, Qin, 2015].

Считается [Auer, Doerge, 2010], что наиболее подходящим методом для проверки нулевой гипотезы и оценки достоверности дифференциальной экспрессии при небольшом размере выборок является точный тест Фишера. Другими тестами, использующимися для

проверки нулевой гипотезы, являются тест отношения правдоподобия, реализованный в пакетах EdgeR [Robinson, McCarthy, Smyth, 2010], DESeq [Love, Anders, Huber, 2013] и DESeq2 [Love, Huber, Anders, 2014], и тест Вальда [Wald, 1945], реализованный в пакетах DESeq и DESeq2. Поскольку оценка дифференциальной экспрессии проводится для всех генов организма, то есть, как правило, для десятков тысяч генов разом, для корректного определения генов, имеющих значимые различия в уровнях экспрессии в разных образцах, необходимо использовать поправки на множественное сравнение. Для этого используются поправка Бонферрони, поправка Бенджамини-Хохберга и другие. Как и в отношении нормализации, обычно в программах для поиска ДЭГ реализованы несколько видов поправок на множественное сравнение, предоставляя выбор пользователю.

На сегодняшний день существует большое число программ, проводящих поиск дифференциальной экспрессии генов. Отдельные исследования были направлены на сравнение качества этих программ и попытки выбрать лучшие из них [Rapaport и др., 2013; Rajkumar и др., 2015; Williams и др., 2017]. В числе лучших программ для поиска дифференциальной экспрессии указывают пакеты для языка R EdgeR, DESeq2, limma-voom. Также, есть упоминания [Williams и др., 2017] о хорошей производительности и точности программы Ballgown, осуществляющей реконструкцию транскриптома из данных картирования. Наконец, в большом исследовании на реальных, то есть не симулированных, данных было показано [Williams и др., 2017], что высокой точностью обладает линейка программ Tuxedo, состоящая из программы TopHat2 для картирования прочтений, программы Cufflinks для подсчёта экспрессии и Cuffdiff для поиска ДЭГ. В то же время, многие авторы критикуют различные программы из этой линейки за низкую точность [Rapaport и др., 2013; Kanitz и др., 2015; Rajkumar и др., 2015].

Оценка качества и эффективности конвейеров программ обычно проводятся на симулированных данных [Kanitz и др., 2015; Kvam, Liu, Yaqing, 2012; Yang и др., 2015]. Программы для симуляции данных экспериментов RNA-seq предоставляют пользователю возможность регулировать параметры создаваемых симулируемых библиотек коротких прочтений [Engström и др., 2013; Frazee и др., 2015; Griebel и др., 2012]. Таким образом, пользователь может исследовать производительность конвейеров и их зависимость от входных данных. Для оценки эффективности поиска дифференциальной экспрессии используется так называемая кривая ошибок (receiver operating characteristic, ROC). Эта

кривая отражает соотношение между количеством истинных положительных результатов, выдаваемых методом, и ложноположительных результатов. Площадь под кривой (area under curve, AUC) показывает точность метода; в данном случае под точность понимается то, насколько эффективно конвейер обнаруживает гены, имеющие дифференциальную экспрессию.

На рисунке 6 приведены отображения на графике кривой ошибок трёх гипотетических конвейеров для поиска дифференциальной экспрессии. «Конвейер_1» при этом представляет некий идеальный случай, когда конвейер определяет как ДЭГ все гены, имеющие в реальности дифференциальную экспрессии, и при этом не включает в множество ДЭГ ни одного гена, не имеющего в реальности дифференциальной экспрессии. Площадь под кривой равна 1. «Конвейер_2», напротив, представляет собой процесс создания списка ДЭГ путём случайного выбора генов независимо от их реальных уровней экспрессии. Площадь под кривой равна 0,5. Для реальных конвейеров программ площадь от кривой лежит в пределах от 0,5 до 1. Чем эта площадь выше, тем более точным и производительным является конвейер [Soneson, Delorenzi, 2013]. Примером такого конвейера является «Конвейер_3», кривая ошибок которого изображена синим цветом на рисунке 6.

При работе с реальными данными исследователю неизвестно, какое количество генов в действительности имеют значимые уровни экспрессии, и какая часть из них значимо изменяет экспрессию. Поэтому нет возможности оценить точность используемых конвейеров с помощью кривой ошибок. Следовательно, требуются другие методы анализа.

В целом, считается, что каждый существующий метод биоинформатической обработки и поиска ДЭГ имеет свои преимущества и недостатки, и производительность отдельных конвейеров зависит от конкретных данных [Conesa и др., 2016]. Для достижения наиболее достоверных результатов рекомендуется проводить поиск ДЭ несколькими методами, после чего берётся пересечение или объединение списков ДЭГ [Rajkumar и др., 2015]. Обычно гены, попадающие в пересечение списков, имеют высокий уровень экспрессии и высокую достоверность ДЭ, в то время как гены, обнаруженные только одним из методов, скорее всего, имеют значения экспрессии и достоверности ДЭ, близкие к пороговым [Lahens и др., 2017]. Наконец, все авторы сходятся во мнении, что чем выше

уровень экспрессии гена хотя бы в одном из образцов, тем точнее будет определена дифференциальная экспрессия.

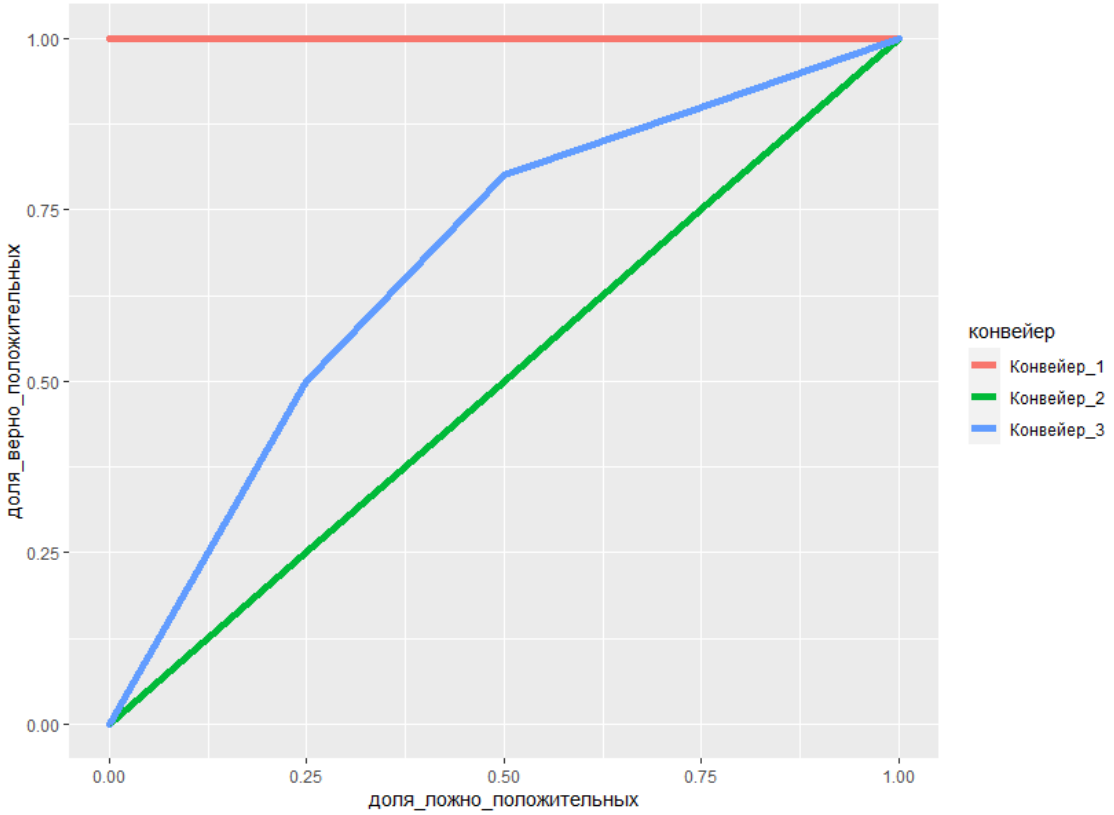

Рис 6. Кривые ошибок для трёх гипотетических конвейеров анализа данных RNA-seq. Площадь под кривыми оператора показывает чувствительность и точность работы конвейеров

Заключение по обзору литературы и постановка задачи исследования

В обзоре литературы были рассмотрены механизмы образования специфических видов окраски колоса у ячменя – частичного альбинизма и частичного меланизма. Показана важность изменения пигментации в жизни растений, влияния на их продуктивность, актуальность изучения этих процессов. Показано, что массовый анализ транскриптома является эффективным методом для изучения подобных механизмов. Рассмотрены современные методы получения и биоинформатического анализ транскриптомных данных у растений. Показано, что несмотря на тщательные методические исследования, проведённые в последние годы, консенсус о наиболее оптимальных методах анализа

данных экспериментов RNA-seq до сих пор не достигнут. Более того, наиболее общепринятое мнение состоит в том, что наиболее оптимальный метод обработки будет специфичным для конкретных данных. Таким образом, учитывая многообразие существующих решений для обработки данных RNA-seq, при анализе конкретных данных необходимо проводить исследование с целью выбора наиболее эффективной стратегии анализа имеющихся данных. Это становится ещё более очевидным, если принять во внимание специфику имеющихся данных, секвенированных на платформе IonTorrent, данные которой отличаются от данных более распространённых платформ Illumina по ряду параметров.

Настоящая работа нацелена на исследование молекулярных механизмов альбинизма и меланизма колоса на основе массового транскриптомного анализа трех линий ячменя – изогенной линии Bowman, которая демонстрирует нормальный фенотип по отношению к окраске колоса, и линий i:Bw*Alm* (частичный альбинизм цветковой чешуи и ряда других органов; Alm – albino lemma) и i:Bw*Blp* (меланизм цветковой чешуи и перикарпа; Blp – Black lemma and pericarp). Такой анализ позволяет изучить гены, вовлечённые в метаболические пути биосинтеза пигментов.

Глава 2. Материалы и методы

2.1 Биологический материал

Для изучения частичного альбинизма ячменя под контролем ядерного гена *Alm*, нуклеотидная последовательность и молекулярные функции которого оставались неизвестными, была использована почти-изогенная линия ячменя i:Bw*Alm* (идентификатор коллекции Nordic GenBank – NGB 20419). Рецессивная мутация в гене *Alm*, локализованном в коротком плече хромосомы 3H ячменя была описана в линии Russia 82 [Takahashi, Hayashi, 1959]. Эта мутация приводит к формированию фенотипа, отличающегося отсутствием хлорофилла в цветковой чешуе, перикарпе, листовом влагалище (первого листа), ушках листовых влагалищ и прилегающей части листовых пластинок, узлах стебля с прилегающей частью междоузлий. Линия i:Bw*Alm* была получена в результате бэккроссирования растений линии Russia 82, имеющих эту мутацию, с растениями сорта Bowman (идентификатор коллекции Nordic GenBank – NGB 22812), с последующим отбором потомства по фенотипу [Nonaka, 1973]. Локус Alm был картирован в коротком плече хромосомы 3H [Lundqvist, Franckowiak, Konishi, 1997]. Сравнительный анализ хромосомы 3H у линии i:Bw*Alm* и Bowman был выполнен Генераловой Галиной Владимировной (лаборант ИЦиГ СО РАН) при использовании микросателлитных маркеров *Xgbms0022, Xgbms0046, Xgbms0048, Xgbms0050, Xgbms0085, Xgbms0102,* Xgbms*0149, Xgbms0212, Xebmac541, Xbmag225*, *Xbmag877*, и продемонстрировал, что сегмент донора-носителя мутантной аллели 3H у линии i:Bw*Alm* ограничен маркерами *Xgbms0050* и *Xgbmag0225* [Shmakov и др., 2016], при этом к короткому плечу хромосомы 3H относится часть этого сегмента, ограниченная маркерами *Xgbms0050* и *Xgbms0149* [Varshney и др., 2007].

В дальнейшем для определённости будем называть участок на коротком плече хромосомы 3H, в котором локализован ген *Alm*, «район Alm». Границы района Alm на физической карте генома ячменя на сегодняшний день не установлены. Однако, было проведено исследование гена *wh* (white husk), локализованного в коротком плече хромосомы 3H, мутация в котором фенотипически проявляется в отсутствии хлорофилла в лемме ячменя и частичном альбинизме других органов растения [Hua и др., 2016]. Район, в котором локализован этот ген, ограничен генами MLOC_14184 и MLOC_80432 с 3’- и 5’-

концов, соответственно [Hua и др., 2016]. Данные идентификаторы генов приведены в аннотации сборки генома линии ячменя Morex, соответствуют идентификаторам генов HORVU3Hr1G031000 и HORVU3Hr1G035380, соответственно, в аннотации сборки генома версии 49.

Из-за сходства локализации генов *Alm* и *wh* (короткое плечо третьей хромосомы) и схожести фенотипического проявления, эти гены рассматриваются как аллельные варианты одного и того же гена [Franckowiak, Lundqvist, Kleinhofs, 2016]. Таким образом, районом Alm, в котором локализован ген *Alm*, будем считать участок на хромосоме 3H, ограниченный генами HORVU3Hr1G031000 и HORVU3Hr1G035380 с 3’- и 5’-концов, соответственно. Этот район содержит 229 генов, аннотированных в текущей сборке генома ячменя. Район Alm, входящие в него гены и расположение на хромосоме 3H, приведены на рисунке 7.

Рис. 7. (А) – хромосома 3H ячменя. Синим цветом отмечены гены малых ядрышковых РНК, красным цветом – район Alm; также отмечены фланкирующие этот район белок-кодирующие гены. (Б) – увеличенный район Alm; отмечены белок-кодирующие гены. (В) – сравнение колосьев ячменя линии i:Bw*Alm* (слева) и сорта Bowman (справа). (Г) – сравнение зёрен ячменя линии i:Bw*Alm* (слева) и сорта Bowman (справа).

Растения линий i:Bw*Alm* и Bowman выращивались в теплице Центра коллективного пользования Лаборатория искуственного выращивания растений ИЦиГ СО РАН. Растения выращивались в условиях 14-часового светового дня и при температурном режиме 20-25ºС. Тотальная РНК была выделена из развивающихся колосков ячменя, включающих цветковую чешую (лемму), по окраске которой отличались i:Bw*Alm* и Bowman. Для выделения РНК в эксперименте с линией i:Bw*Alm* был использован набор Plant RNA MiniPrepTM kit (Zymo Research Corporation, Ирвин, Калифорния, США); в эксперименте с линией BLP был использован набор RNeasy Plant Mini Kit (QIAGEN, Германия). Выделение РНК проводилось Генераловой Галиной Владимировной и Кукоевой Татьяной Владимировной (ИЦиГ СО РАН). Для каждой из двух линий было приготовлено по три биологических повторности. Для получения каждой повторности биологический материал трёх разных растений, относящихся к этой линии, был объединён, после чего было проведено выделение тотальной РНК.

Другая линия, исследованная в данной работе – почти изогенная линия i:Bw*Blp* (идентификатор коллекции Nordic GenBank – 20470). Растения этой линии имеют чёрную меланин-подобную окраску цветковой чешуи и перикарпа. Такой фенотип встречается в некоторых популяциях дикорастущего и культурного ячменя. Данная линия получена путём возвратных скрещиваний черноколосого ячменя с неокрашенным сортом Bowman с отбором потомства по фенотипу [Buckley, 1930]. Ген *Blp*, доминантный аллель которого вызывает формирование описанного фенотипа, находится в длинном плече хромосомы ячменя 1H [Robertson и др., 1965].

По аналогии с районом Alm, будем называть участок на длинном плече хромосомы 1H, в котором локализован ген *Blp*, «район Blp». Позднее ген *Blp* был локализован на физической карте хромосом ячменя с большей точностью – он расположен на хромосоме 1H между генами HORVU1Hr1G086690 и HORVU1Hr1G087070 [Long и др., 2019]. В текущей сборке генома ячменя в данном участке хромосомы 1H локализован 21 ген. Район Blp показан на рисунке 8.

Растения были выращены в теплице Центра коллективного пользования Лаборатория искусственного выращивания растений ИЦиГ СО РАН. Растения выращивались в условиях 12-часового светового дня и при температурном режиме 20-25ºС. Тотальная РНК была выделена из цветковой чешуи (леммы) с перикарпом растений ячменя

линий i:Bw*Blp* и Bowman во время ранней фазы восковой спелости. Для выделения РНК в эксперименте с линией i:Bw*Blp* был использован набор RNeasy Plant Mini Kit (QIAGEN, Германия). Выделение РНК проводилось Шоевой Олесей Юрьевной и Глаголевой Анастасией Юрьевной (ИЦиГ СО РАН). Тотальная РНК, выделенная из нескольких растений, была объединена для получения каждого из биологических образцов. Так было приготовлено по три биологических образца в исследуемой почти изогенной линии и контрольной изогенной линии. Далее было проведено обогащение поли-А-фракции РНК путём инкубации образцов с поли-Т-содержащими гранулами, после чего РНК была фрагментирована инкубацией в присутствии эндонуклеаз.

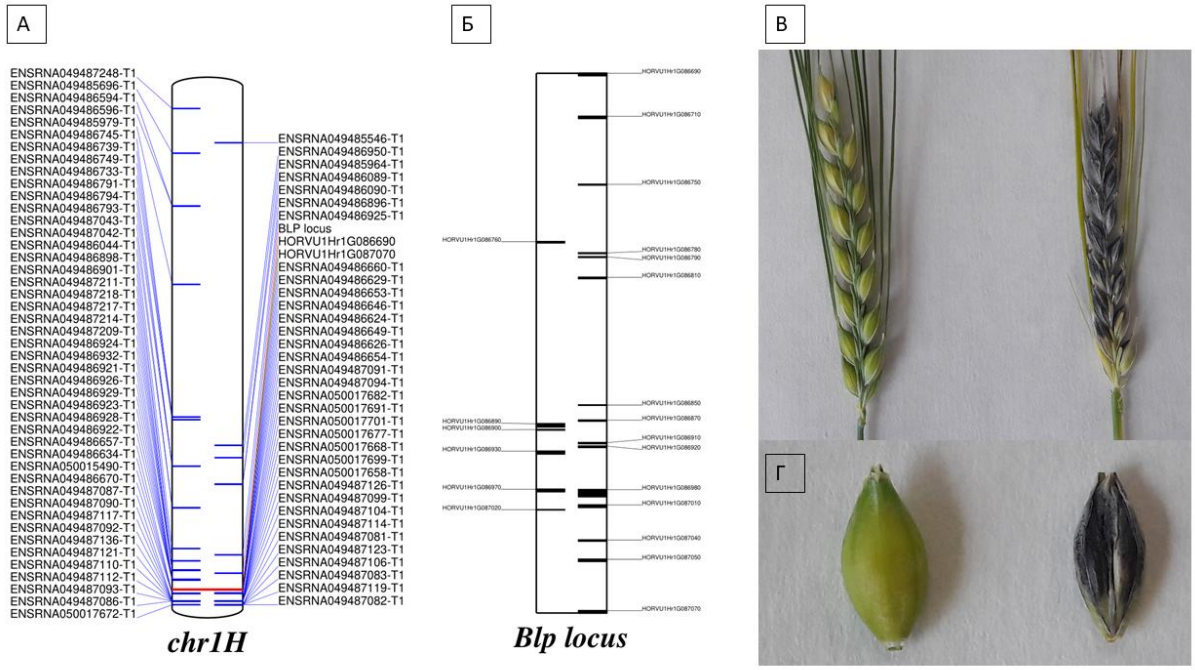

Рис. 8. (А) – хромосома 1H ячменя. Синим цветом отмечены гены малых ядрышковых РНК, красным цветом – район Blp; также отмечены фланкирующие этот район белок-кодирующие гены. (Б) – увеличенный район Blp; отмечены белок-кодирующие гены. (В) – сравнение колосьев ячменя сорта Bowman (слева) и линии i:Bw*Blp* (справа). (Г) – сравнение зёрен ячменя сорта Bowman (слева) и линии i:Bw*Blp* (справа).

Библиотеки коротких прочтений, полученные в двух экспериментах, согласно процедурам, описанным выше, далее были секвенированы Васильевым Геннадием

Владимировичем и Шацкой Натальей Владиславовной (ИЦиГ СО РАН) Для секвенирования была использована платформа Ion Torrent PGM.

Верификация дифференциальной экспрессии была проведена с помощью количественной полимеразной цепной реакции в реальном времени. Предварительно образцы РНК были обработаны ДНКазами (QIAGEN RNase-Free DNase Set). Набор RevertAid™ kit (Thermo Fisher Scientific Inc., Waltham, Массачусетс, США) был использован для приготовления библиотек клонов ДНК на основе используемой РНК. Для проведения ПЦР был использован набор SYNTOL SYBR Green I (Синтол, Москва, Россия). В качестве контроля была взята экспрессия гена убиквитина *Ubc*. Для дизайна пар праймеров для ПЦР был использован онлайн-сервис IDT PrimerQuest (http://eu.idtdna.com/PrimerQuest/Home/). Списки генов, экспрессия которых была верифицирована с помощью ПЦР, приведены в дополнительной таблице 1. Количественная ПЦР была проведена Глаголевой Анастасией Юрьевной (ИЦиГ СО РАН).

Семена изучаемых почти изогенных линий и сорта Bowman были любезно предоставлены Нордическим генбанком (Nordgen, nordgen.com, Швеция). В классификации этого банка семян, использованные линии имеют идентификаторы NGB 22812 (Bowman), NGB 20470 (BLP) и NBG 20419 (i:Bw*Alm*).

Для хромосомной локализации *de novo* реконструированного транскрипта использовали пшенично-ячменные дополненные линии Chinese Spring / Betzes [Hart, Islam, Shepherd, 1980], ДНК которых любезно предоставила доктор Марион Рёдер (IPK-Гатерслебен, Германия). Процедура локализации контига в геноме ячменя вплоть до хромосомного плеча была проведена Глаголевой Анастасией Юрьевной.

2.2 Биоинформатический анализ библиотек коротких прочтений

Как отмечается в литературе, существует множество программных конвейеров обработки данных экспериментов RNA-seq, и в каждом конкретном случае необходим подбор конвейера, который окажется наиболее оптимальными для имеющихся данных [Conesa и др., 2016; Yang и др., 2015]. Разные существующие виды обработки более подходят для разных исходных данных.

Таким образом, для максимальной эффективности обработки экспериментов RNA-seq требуется сперва оценить, насколько разные существующие методы подходят для

анализа эти данных. Необходимо также учитывать, что биоинформатическая обработка состоит из нескольких стадий – чаще всего, фильтрации сырых библиотек, картирования, подсчёта и нормализации уровней экспрессии генов и, наконец, поиска дифференциальной экспрессии. Для проведения каждой из этих стадий существует множество программных продуктов. Таким образом, выбор оптимального подхода для обработки может быть нетривиальным решением.

Для разрешения этой проблемы обработка имеющихся данных была проведена с использованием разных программ на каждой стадии обработки. Этот процесс схематично показан на рисунке 9 и более подробно рассмотрен в следующих разделах.

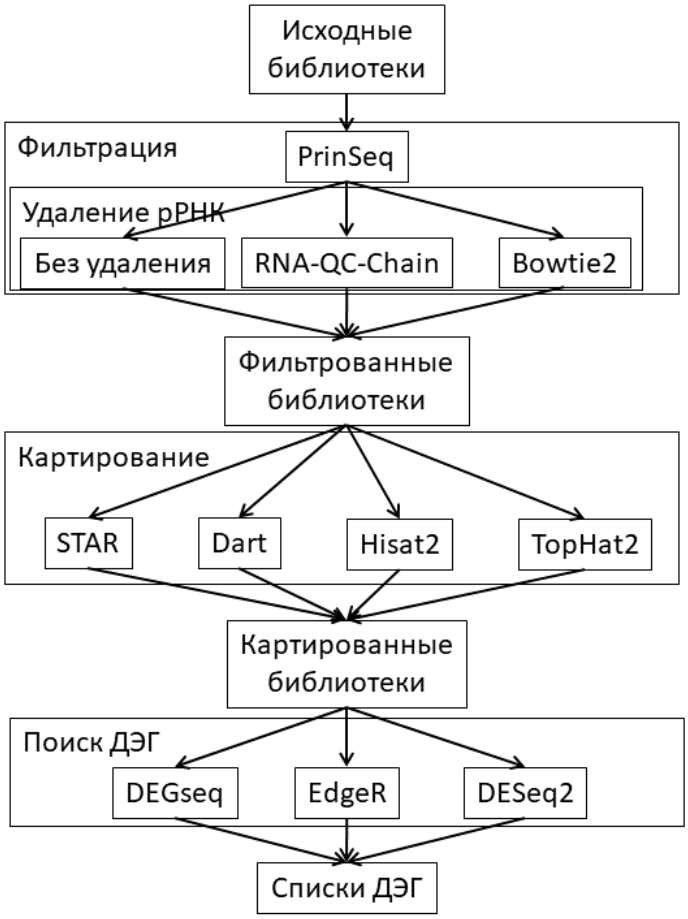

Рисунок 9. Конвейер биоинформатической обработки библиотек коротких прочтений, использованных в данной работе. Показаны три основные стадии обработки данных – фильтрация, картирование и поиск ДЭГ, и приведены использованные на каждой из стадий программы.

2.2.1 Фильтрация библиотек

Для оценки качества библиотек коротких прочтений была использована программа FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) версии 0.11.5. Далее, была использована программа Cutadapt [Martin, 2011] версии 1.9.1 для удаления встречающихся в прочтениях адаптерных последовательностей. Программа Prinseq-lite [Schmieder, Edwards, 2011] версии 0.20.4 была использована для фильтрации библиотек по длинам и качеству прочтений: были удалены прочтения со средним значением качества Phred 20 или ниже, и прочтения с длинами меньше 50 нуклеотидов или больше 270 нуклеотидов. Выбранные ограничения по длинам прочтений связаны с тем, что прочтения длинами меньше 50 нуклеотидов могут некорректно восприниматься программами для картирования; в то же время, как можно видеть из оценки качества прочтений, для большинства прочтений с длинами больше 270 качество секвенирования резко падало (см. результаты, раздел «предобработка библиотек»).

При оценке качества библиотек с помощью программы FastQC были отдельно рассмотрены перепредставленные последовательности. Обнаруженные последовательности, встречающиеся не менее чем в 1% прочтений хотя бы в одной библиотеке были выровнены на нуклеотидные последовательности базы данных NCBI Nucleotide с помощью онлайн-сервиса Blastn.

Для удаления последовательностей рРНК из библиотек были использована программа RNA-QC-Chain [Zhou и др., 2018] версии 1.0. В качестве альтернативного подхода к удалению последовательностей рРНК предлагается картировать библиотеки на последовательности рРНК, после чего брать для дальнейшего анализа те прочтения, которые не были картированы [Nguyen и др., 2014]. Для этого была использована программа Bowtie2 [Langmead, Salzberg, 2012] версии 2.3.4. Последовательности некодирующих РНК ячменя были получены из базы данных Ensembl Plants [Kersey и др., 2018], после чего последовательности были индексированны программой bowtie2-build. Далее, библиотеки были выровнены на последовательности рРНК, и прочтения, которые не были успешно картированы, были взяты в дальнейшую обработку. Наконец, чтобы оценить влияние фильтрации рРНК на качество последующей обработки библиотек, было решено также использовать в дальнейшем анализе библиотеки, не прошедшие удаление рРНК, и

провести сравнение качества работы с такими библиотеками и с библиотеками, очищенными от рРНК.

Таким образом, были использовано три подхода к фильтрации библиотек – без удаления рРНК, с удалением рРНК с помощью RNA-QC-Chain и с удалением рРНК с помощью Bowtie2.

2.2.2 Картирование библиотек и подсчёт уровней экспрессии

Картирование библиотек на референсный геном было проводено четырьмя программами: STAR [Dobin и др., 2013] версии 2.6.1a, Dart [Lin, Hsu, 2018a] версии 1.3.2, HISAT2 [Kim, Langmead, Salzberg, 2015] версии 2.1.0, TopHat2 [Kim и др., 2013a] версии 2.1.1. В качестве референса для картирования прочтений библиотек была использована сборка генома ячменя, находящаяся в свободном доступе в базе данных Ensembl plants [Kersey и др., 2018]. Отметим, что эта сборка генома ячменя была проведена на материале изогенной линии Morex [Mascher и др., 2017], в то время как данная работа проводится на РНК изогенной линии Bowman и почти изогенных линий, полученных на основе линии Bowman. Геном ячменя линии Bowman на сегодняшний день собран до состояния контигов, часть из которых привязана к отдельным хромосомам или плечам хромосом, однако не составлена в единую карту, на которой было бы отмечено взаимное расположение контигов. Таким образом, в силу того, что сборка генома линии Morex является на сегодняшний день более полной, она была использована в качестве референса для картирования.

Последовательности генома ячменя линии Morex в формате fasta, разметка генома в формате gff3 были скачаны из базы данных Ensembl plants, и было проведено индексирование генома каждой из программ: STAR с использованием опции –runMode genomeGenerate, dart с использованием скрипта bwt_index, hisat2 с помощью скрипта hisat2-build, TopHat2 с помощью скрипта bowtie_build. С помощью каждой из программ библиотеки, полученные в результате каждого из трёх подходов к фильтрации, были картированы на референсный геном ячменя. Итого, для каждой библиотеки были получены 12 картирований.

Для подсчёта уровней экспрессии была использована программа FeatureCounts из пакета Subread [Liao, Smyth, Shi, 2014]. Подсчёт экспрессии производился на уровне

экзонов, с использованием опции ‘–feature’. В результате были получены таблицы, отражающие количество прочтений, картированных на каждый из генов ячменя – так называемые «матрицы экспрессии». Эти таблицы были далее использованы для поиска дифференциальной экспрессии генов.

Таким образом, на стадии картирования библиотек и подсчёта экспрессии генов были использованы три подхода, что, в комбинации с тремя подходами к фильтрации библиотек, даёт двенадцать различных способов обработки данных.

2.2.3 Поиск дифференциальной экспрессии генов

Поиск ДЭГ осуществлялся тремя пакетами для R: edgeR [Robinson, McCarthy, Smyth, 2010], DEGseq [Wang и др., 2009] и DESeq2 [Love, Huber, Anders, 2014]. Предварительно для всех генов было подсчитано значение cpm (counts per million). После этого все гены, не имеющие значения cpm больше двух хотя бы в двух образцах из шести, были удалены из дальнейшего рассмотрения.

Для поиска ДЭГ пакетом EdgeR было проведён подсчёт дисперсии внутри групп, соответствующих генотипам i:Bw*Alm* и Bowman, и между групп, после чего с помощью точного теста Фишера была подсчитана достоверность дифференциальной экспрессии каждого гена между линиями i:Bw*Alm* и Bowman. Была проведена поправка на множественное сравнение по методу Бенджамини-Хохберга.

Для поиска ДЭГ пактом DESeq2 был проведён подсчёт дисперсии экспрессии генов между группами и внутри групп, после чего было проведено моделирование экспрессии генов отрицательным биномиальным распределением, и проведён тест отношения правдоподобия для подсчёта достоверности дифференциальной экспрессии гена между группами. Поправка на множественное сравнение была проведена с помощью метода Бенджамини-Хохберга.

При использовании пакета DEGseq был использован метод MARS [Wang и др., 2009], являющийся адаптацией метода, основанного на MA-графиках, использованного ранее для анализа данных микрочипов [Yang и др., 2002]. Достоверность дифференциальной экспрессии для каждого гена была подсчитана с помощью теста отношения правдоподобия, после чего применена поправка Бенджамини-Хохберга на множественное сравнение.

2.2.4 Резюме биоинформатической обработки

Для биоинформатического анализа данных RNA-seq, связанных с процедурой картирования, были опробованы разные подходы, состоящие в комбинировании использованных на разных стадиях анализа программных средств. Так, было использовано три подхода для фильтрации библиотек – фильтрация без удаления последовательностей рРНК, фильтрация с удалением рРНК путём картирования на последовательности рРНК ячменя, и фильтрация с удалением рРНК с помощью программы RNA-QC-Chain. Далее, на стадии картирования были использованы программы Dart, Hisat2, Star и TopHat2. Наконец, для поиска дифференциальной экспрессии генов были использованы пакеты DEGseq, DESeq2 и EdgeR. Таким образом, было опробовано в общей сложности 36 конвейеров программ.

Для сравнения производительности конвейеров и поиска конвейера, оптимального для анализа имеющихся данных RNA-seq в каждом из экспериментов, для каждого из 36 использованных конвейеров были оценены четыре показателя производительности. Первые два показателя – это доля всех картированных на геном ячменя прочтений, и доля уникально картированных прочтений. Чем эти показатели выше, тем больше точность картирования библиотек данным конвейером. Эти показатели будем далее называть ‘%_mapped’ и ‘%_uniq’.

Третий показатель, взятый для оценки качества конвейеров, иллюстрирует точность определения ДЭГ. Для количественной оценки точности поиска ДЭГ в каждому из двух экспериментов (с линией i:Bw*Alm* и с линией i:Bw*blp*) была составлена выборка генов, для которой была проведена верификация дифференциальной экспрессии с помощью количественной ПЦР в реальном времени. Далее был подсчитан коэффициент корреляции Пирсона между определёнными путём анализа RNA-seq и наблюдаемыми экспериментально с помощью количественной ПЦР значениями изменения экспрессии данного списка верифицированных генов. Чем коэффициент корреляции выше, тем более точно данный конвейер определяет ДЭГ. Этот показатель далее будем называть ‘corr’.

Наконец, была оценена стабильность точного определения ДЭГ. Оценка стабильности точного определения ДЭГ конвейером проводилась следующим образом. Из списка генов, для которых проводилась верификация дифференциальной экспрессии с помощью количественной ПЦР в каждом из экспериментов, было составлено *n* подвыборок

размером *n-1*, где *n* – количество генов, для которых была проведена верификация. Далее, был подсчитан коэффициент корреляции Пирсона между идентифицированными в ходе анализа RNA-seq и полученными экспериментально значениями изменения экспрессии, для каждой из полученных подвыборок генов. После этого для полученных значений корреляции для этого конвейера было подсчитано значение среднеквадратичного отклонения. Это значение характеризует стабильность точного определения ДЭГ с помощью данного конвейера. Чем это значение ниже, тем более стабильно конвейер определяет изменение экспрессии. Этот показатель далее будем называть ‘stdev’.

В таблице 1 приведены примеры значений данных показателей для трёх отдельно взятых конвейеров. Отметим, что таблица 1 служит лишь для иллюстрации общего принципа приоретизации конвейеров, и приведённые в ней данные не используются в дальнейшем в работе.

Таблица 1. Конвейеры биоинформатической обработки и количественные показатели, использованные для их приоретизации

Входящие в конвейеры программы			Показатели			
Метод фильтрации	Программа для картирования	Пакет для поиска ДЭГ	corr	stdev	%_mapped	%_uniq
bowtie2	Dart	EdgeR	0,8382	0,0579	98,77	87,40
RNA-QC-Chain	Star	DEGseq	0,8070	0,0562	98,70	79,05
bowtie2	Hisat2	DESeq2	0,7986	0,0675	72,07	62,06

По каждому из этих четырёх параметров конвейеры были ранжированы, и затем была подсчитанна сумма рангов. Конвейер, для которого такая сумма рангов оказалась максимальной, использованных для этого эксперимента, выбирался как наиболее оптимальный для работы с имеющимися данными RNA-seq. В случае, если несколько конвейеров имеют одинаковую сумму рангов, предпочтения отдаётся тому конвейеру, который имеет наибольшее значение коэффициента корреляции Пирсона с результатами количественной ПЦР для контрольных генов.

Для трёх конвейеров, приведённых в качестве примера в таблице 2, порядковые номера после сортировки по каждому из показателей и полученная сумма порядковых номеров приведены в таблице 2.

Следует отметить, что показатели ‘%_mapped’ и ‘%_uniq’ различаются только между конвейерами, в которых были использованы разные подходы к фильтрации библиотек и разные программы для картирования. У конвейеров же, различающихся только использованным пакетом для поиска ДЭГ, эти значения одинаковы. Таким образом, при сортировке по каждому из этих двух показателей, конвейерам можно присвоить порядковые значения только от 1 до 12. В то же время, при сортировке по показателям ‘corr’ и ‘stdev’ конвейерам присваиваются значения от 1 до 36. Таким образом, показатели ‘%_mapped’ и ‘%_uniq’ вносят меньший вклад в общую сумму показателей для каждого из конвейеров.

Табл. 2. Приоретизация конвейеров биоинформатической обработки на основании суммы рангов использованных количественных показателей

Составляющие части конвейера			Ранги				Сумма рангов
Метод фильт-рации	Программа для картиро-вания	Пакет для поиска ДЭГ	corr	stdev	%_mapped	%_uniq	
bowtie2	Dart	EdgeR	3	2	3	3	11
RNA-QC-Chain	Star	DEGseq	2	3	2	2	9
bowtie2	Hisat2	DESeq2	1	1	1	1	4

С целью определить вклад каждой из переменных в качество дальнейшего анализа был использован метод главных компонент для имеющихся 36 конвейеров биоинформатической обработки. В качестве переменных выступали приведённые выше четыре рассмотренных показателя качества работы конвейера. Для анализа главных компонент использована команда prcomp языка программирования R, при этом все переменные были нормированы и приведены к одному значению стандартного отклонения,

для чего был задан параметр ‘scaling=T’. С помощью пакета ggplot2 были построены графики, на которых конвейеры биоинформатической обработки распределены по первым двум главным компонентам.

2.2.5 Функциональный анализ ДЭГ

Для генов, имеющих дифференциальную экспрессию в каждом из экспериментов (с мутантами Alm и Blp) был проведен поиск значимо ассоциированных терминов онтологий и метаболических путей. Для поиска представленных терминов ГО списки ДЭГ были обработаны с помощью онлайн-сервиса Singular Enrichment Analysis (SEA), предоставленного базой данных AgriGO v2 [Tian и др., 2017]. В качестве референса был использован набор идентификаторов генов IPK Barley BLAST Server. Для подсчёта достоверности был использован гипергеометрический тест с последующей поправкой по методу Бенджамини-Хохберга. Гены с повышенной и пониженной экспрессией были оценены отдельно.

Для анализа участия генов в известных метаболических путях была использована база данных PlantCyc [Schläpfer и др., 2017]. Имеющиеся в базе данных списки генов, участвующих в метаболических путях, были скачаны на локальный компьютер. После этого с помощью скриптов на языках Perl и R был проведён анализ обогащения метаболических путей следующим образом: для каждого метаболического пути было подсчитано количество генов, участвующих в нём, и было определено, какое количество из этих генов являются ДЭГ в изучаемом эксперименте. Также было подсчитано общее количество генов ячменя, упомянутых в базе данных PlantCyc как участвующие в тех или иных метаболических путях. Наконец, известно общее количество аннотированных в текущей сборке генома генов ячменя, и известно общее количество ДЭГ. Далее с помощью функции hypergeom языка R для каждого метаболического пути было подсчитано значение статистического критерия *p*, показывающего вероятность получения такого количества ДЭГ среди генов, входящих в данный метаболический путь при их случайном распределении, подчиняющемся гипергеометрическому распределению:

$$P(x) = \frac{\binom{m}{x}\binom{M-m}{k-x}}{\binom{M}{k}} \qquad (12)$$

Здесь x – количество ДЭГ, входящих в данный метаболический путь; m – количество генов, входящих в данный метаболический путь; M – количество генов ячменя, входящих в базу данных PlantCyc; k – количество ДЭГ в данном эксперименте. Как и в случае с анализом терминов ГО, обработка генов с повышенной и пониженной экспрессией проводилась по отдельности. Далее, с помощью функции языка R p.adjust была сделана поправка на множественное сравнение по методу Бенджамини-Хохберга.

Отдельно был произведён анализ изменения экспрессии генов, входящих в районы Alm и Blp, соответственно. Необходимо отметить, что, несмотря на усилия мирового сообщества по изучению генетики ячменя, существующая аннотация генома является неполной. Так, только для 8885 из 38029 белок-кодирующих генов, представленных в аннотации версии IBSCv2 генома ячменя, приведено описание белкового продукта. Для определения функции остальных генов была рассмотрена доступная в базе данных Ensembl plants информация о доменной структуре их белковых продуктов, а также ассоциированные вхождения в других базах данных, содержащих информацию о белковой структуре: Pfam [Mistry и др., 2020], PANTHER [Thomas и др., 2022], TIGFAM [Haft и др., 2013] и Superfamily [Gough и др., 2001].

Наконец, была рассмотрена экспрессия генов, локализованных в пластоме. Они были подразделены на три категории в соответствии со своими функциями: гены, кодирующие белки, входящие в состав фотосистем или участвующие в фотосинтезе иным способом; гены, кодирующие рибосомные белки; проче гены. Списки генов пластома, входящие в каждую из этих категорий, приведены в дополнительной таблице 2. Далее, для каждой из трёх категорий было подсчитано среднее значение изменения экспрессии всех входящих в неё генов, и среднеквадратичное отклонение этого параметра. Затем была подсчитана значимость наблюдаемых различий в изменениях уровней экспрессии для разных групп с помощью теста Манна-Уитни.

2.2.6 *de novo* реконструкция транскриптома

Прежде всего, для проведения сборки транскриптома *de novo* требовалось установить, какой из способов фильтрации библиотек коротких прочтений является наиболее оптимальным, чтобы в дальнейшем использовать очищенные библиотеки, полученные этим способом, для проведения *de novo* сборки. Процедура выбора

оптимального конвейера обработки, описанная в разделе «методы: биоинформатический анализ библиотек коротких прочтений», включает в себя выбор наиболее эффективной процедуры фильтрации библиотек. Таким образом, после проведения биоинформатической обработки библиотек и выделения оптимального конвейера способ фильтрации библиотек, входящий в этот конвейер, был установлен как оптимальный способ фильтрации этих библиотек. Соответственно, библиотеки, прошедшие данный способ фильтрации, были использованы для проведения *de novo* сборки транскриптома ячменя.

Была проведена сборка транскриптома ячменя *H. vulgare* почти изогенной линии i:Bw*Alm* с частичным альбинизмом колоса и стебля с помощью трёх программ для реконструкции транскриптома *de novo*: Trinity [Grabherr и др., 2013] версии 2.2.0, Trans-ABySS [Robertson и др., 2010] версии 2.0.1 и Spades [Bushmanova и др., 2018] версии 3.12.0. Схема анализа подробнее показана на рисунке 10.

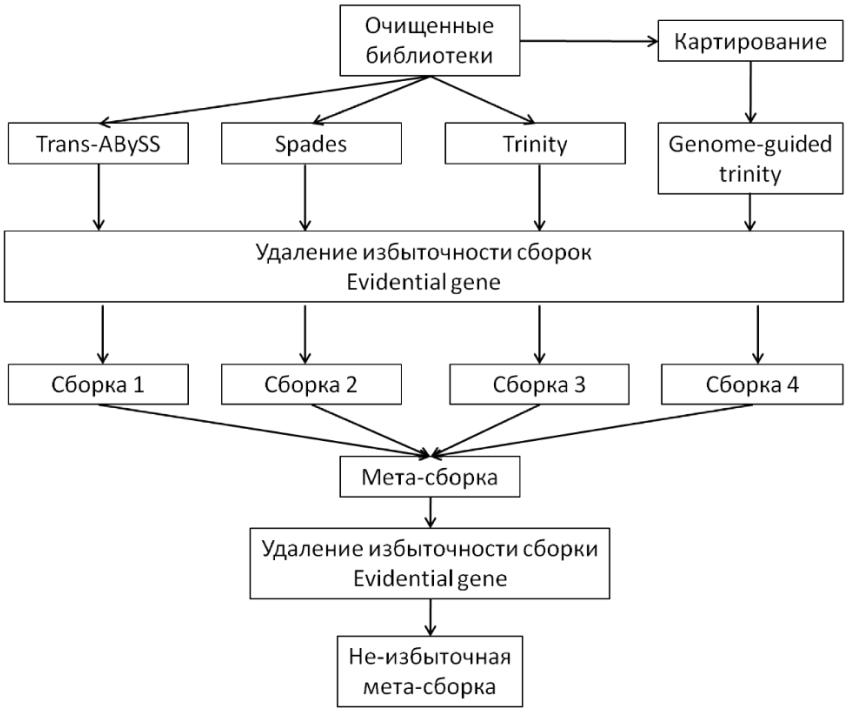

Рис 10. Схема получения мета-сборки *de novo* транскриптома леммы ячменя изучаемых изогенных линий i:Bw*Alm* и i:Bw*Blp*

Запуск сборщика Trinity проходил с параметрами «по умолчанию», на ввод программы были поданы шесть библиотек, относящихся к данному эксперименту. При

запуске программы Spades на ввод также были поданы все шесть библиотек коротких прочтений, относящиеся к данному эксперименту. При запуске программы Spades были указаны опции ‘--iontorrent’ и ‘--only-assembler’.

Сборка программой Trans-ABySS была проведена для каждой из библиотек, относящихся к данному эксперименту, по отдельности, после чего программой transabyss-merge, входящей в пакет программ Trans-ABySS, полученные сборки были объединены. Эта сборка проходила с параметрами «по умолчанию», при которых длина k-мера равна 32. Аналогичным образом были проведены сборки со значениями параметра k 48 и 64. Таким образом, с помощью Trans-ABySS были получены три сборки *de novo*, отличающиеся длинами k-меров, использованными при работе. После этого три полученных сборки были объединены программой transabyss-merge. Результирующая сборка была далее использована как индивидуальная сборка транскриптома *de novo*, полученная с помощью программы trans-ABySS.

Дополнительно, была проведена геном-ориентированная сборка программой Trinity. Для этого, сперва из 12 имеющихся картирований библиотек было выбрано наиболее оптимальное картирование (см. раздел «Методы: биоинформатический анализ»). Затем из файлов картирования библиотек в формате sam (sequence alignment/mapping) был получен общий файл, объединяющий все шесть картирований, при помощи команды merge программы samtools версии 1.6. Этот файл, вместе с шестью библиотеками, относящимися к данному эксперименту, был использован для сборки программой Trinity в режиме геном-ориентированной сборки транскриптома, с указанием при этом максимальной длины интрона равной 500000 нуклеотидов.

Важная стадия получения высококачественной сборки транскриптома *de novo* – удаление избыточности сборки, то есть удаление каждого контига, который является подсловом хотя бы одного другого контига, входящего в данную сборку. Для более подробного ознакомления см. раздел 1.6.3.3. Для удаления избыточности сборок была задействована программа tr2aacds.pl из линейки программ Evidential Gene версии 20.05.2020 г. [Gilbert, 2019]. Каждая из сборок была обработана этой программой по отдельности. Таким образом, были получены три не-избыточные сборки транскриптома *de novo* и одна не-избыточная геном-ориентированная сборка транскриптома. В дальнейшем для простоты будем называть *de novo* сборки, полученные разными программами,

сокращёнными называниями этих программ: abyss, spades и trinity для сборок, полученных с помощью Trans-ABySS, RNA-Spades и Trinity, соответственно. Геном-ориентированную сборку будем далее называть сокращённо GG (от англ. Genome-Guided – геном-ориентированная). На рисунке 10 изображена схема, иллюстрирующая основные этапы первого подхода получения не-избыточной мета-сборки.

Для получения оптимального мета-транскриптома сборки были конкатенированы в один файл, после чего этот файл для удаления избыточности также был обработан программой tr2aacds.pl. Следует отметить, что здесь и далее рассматриваются контиги, имеющие открытые рамки считывания, так как tr2aacds.pl использует для дальнейшего анализа только контиги, в которых были идентифицированы открытые рамки считывания, имеющие длину не меньше пороговой.

Таким образом, для каждого из двух экспериментов было получено по четыре индивидуальных сборки транскриптома: сборки spades и trinity, составленные каждая из всех шести библиотек коротких прочтений, входящих в этот эксперимент; сборка abyss, проведённая для каждой из библиотек по отдельности с разными значениями k-меров, после чего сборки для разных библиотек были объединены в одну сборку abyss с помощью программы abyss-merge; геном-ориентированная сборка GG, составленная из всех шести библиотек, входящих в этот эксперимент, и файла картирования, объединённого из файлов картирования всех шести библиотек, входящих в этот эксперимент, на геном ячменя. Далее из четырёх индивидуальных сборок для каждого из экспериментов была получена одна мета-сборка транскриптома ячменя

2.2.7 Анализ de novo сборки транскриптома

Для оценки качества реконструкции транскриптома каждой из почти изогенных линий все отдельные сборки и мета-сборка транскриптома этой линии прошли обработку следующими программами: BUSCO версии 3.0.2 для оценки полноты сборок на основании представленности характерных для растений последовательностей; Transrate версии 1.0.3 для аннотации контигов и оценки полноты наличия генов ячменя в сборке. После этого было проведено сравнение наборов CDS ячменя, обнаруженных программой Transrate в каждой из индивидуальных сборок. На основании перекрывания множеств, обнаруженных в каждой из индивидуальных сборок CDS, были построены диаграммы Венна,

иллюстрирующие вклад каждого из сборщиков транскриптома *de novo* в структуру мета-сборки.

Далее контиги двух мета-сборок транскриптома ячменя, относящиеся к двум экспериментам, были выравнены на последовательность генома ячменя *H. vulgare*, с помощью программы rnaQUAST [Bushmanova и др., 2016]. rnaQUAST подсчитывает и предоставляет для оценки пользователя различные параметры, полученные из выравнивания контигов на референс, благодаря чему можно оценить качество сборки. В частности, эта программа разделяет контиги на три категории: контиги, выровненные на референс и совпадающие с аннотированными генами; контиги, выровненные на референс, но не совпадающие с известными аннотированными генами; и контиги, не имеющие существенной гомологии к референсному геному. Эту последнюю группу контигов будем далее называть «новыми контигами».

Сравнение качества сборок транскриптома

Для количественного сравнения качества сборок был использован подход, предложенный в работе Хольцера и Марца [Hölzer, Marz, 2019]. Он состоит в том, чтобы для ряда выбранных параметров, отражающих качество сборки транскриптома de novo, провести процедуру нормализации по следующей формуле:

$$N_j^i = \frac{R_j^i - \min(V^i)}{\max(V^i) - \min(V^i)} \quad (13)$$

Где R_j^i – значение параметра i для сборки транскриптома j до нормализации, N_j^i – значение этого параметра после нормализации, V^i – вектор, составленный из всех значений параметра i для всех сборок транскриптома de novo до нормализации:$V^i = (V_1^i, \dots, V_5^i)$. Таким образом, после нормализации каждый из параметров принимает значение от 0 до 1 для каждой из сборок de novo. После этого для каждой из сборок все значения нормализованных параметров суммируются, после чего проводится градация сборок по значению суммы всех нормализованных параметров. Сборка, имеющая наибольшую сумму нормализованных параметров, считается наиболее качественной.

Для сравнения качества индивидуальных сборок и мета-сборок транскриптома ячменя, полученных при работе с библиотеками коротких прочтений, относящихся к двум экспериментам, были использованы следующие семь параметров, характеризующих

разные аспекты качества сборки транскриптома: (1) N50; (2) медиана распределения длин контигов; (3) количество обнаруженных генов из списка BUSCO (как обнаруженных целиком, так и фрагментарно); (4) доля контигов, для которых была обнаружена гомология с известными CDS ячменя с помощью TransRate; (5) количество CDS ячменя, с которыми контиги из сборки de novo имеют гомологию; (6) количество CDS ячменя, не менее 95% длины которых покрыто выравниванем с контигами из сборки *de novo*; (7) доля прочтений из библиотек, использованных для создания сборки *de novo*, псевдо-картированных на эту сборку с помощью программы kallisto.

Характеристики 1 и 2 отражают распределение длин контигов. Характеристики 3, 4, 5 и 6 отражают полноту сборки транскриптома. Характеристика 7 отражает полноту сборки транскриптома и полноту использования библиотек коротких прочтений при составлении этой сборки.

Полученные новые контиги были проанализированы более подробно. Аминокислотные последовательности, кодированные в ОРФ, обнаруженных в транскриптах, были выровнены на базу данных NCBI Protein nr с помощью онлайн-сервиса blastn [Camacho и др., 2009]. Контиги, чьи аминокислотные продукты не имеют значимой гомологии с содержащимися в базе данных белковыми последовательностями, были удалены из дальнейшего рассмотрения как артефакты сборки *de novo*. Контиги, аминокислотные продукты которых имеют наибольшую гомологию с последовательностями, не относящимися к растительным (*Viridiplantae*), были удалены из дальнейшего рассмотрения как контаминация чужеродным генетическим материалом.

Из оставшихся новых контигов были выделены те, для которых была показана достоверная дифференциальная экспрессия. Для этих контигов была проведена оценка кодирующего потенциала с помощью онлайн-сервиса Coding Potential Calculator 2 [Kang и др., 2017]. Этот сервис оценивает длину открытой рамки считывания и кодируемого аминокислотного продукта в сравнении с известными распределениями таких длин для белок-кодирующих транскриптов; также оцениваются изоэлектрическая точка аминокислотного продукта, целостность ОРФ и значение критерия Фикетта [Fickett, 1982].

Далее, для этих контигов были рассмотрены 100 лучших гомологов их аминокислотных продуктов из базы данных NCBI Protein nr. Среди гомологов были проанализированы только имеющие функциональную аннотацию. Также была рассмотрена

доменная структура, обнаруженная в аминокислотных продуктах с помощью онлайн-сервиса поиска консервативных доменов базы данных NCBI Structure [Lu и др., 2020]. Наконец, был проведён поиск этих аминокислотных продуктов по базе данных InterPro [Paysan-Lafosse и др., 2023] с помощью онлайн-сервиса InterProScan [Jones и др., 2014] для определения их доменной структуры.

Глава 3. Результаты

3.1 Анализ транскриптома почти изогенной линии ячменя i:Bw*Alm* в сравнении с Bowman

Был проведён биоинформатический анализ библиотек коротких прочтений, полученных в эксперименте с линией ячменя i:Bw*Alm*. Почти изогенная линия ячменя i:Bw*Alm* отличается дефицитом хлорофилла в лемме, ушках, стебле и колосковой чешуе. Формирование данного фенотипа обусловлено мутацией в гене *Alm*, локализованном в коротком плече хромосомы 3H. На данный момент структура и функции гена *Alm* неизвестны. Сравнительный анализ транскриптома леммы ячменя линии i:Bw*Alm* и сорта Bowman, имеющего нормальный фенотип по окраске колоса, может помочь выявить гены, участвующие в формировании данного фенотипа. Это, в свою очередь может пролить свет на особенности метаболизма хлорофилла и контроль этого процесса у злаков.

3.1.1 Предобработка библиотек коротких прочтений

В работе были использованы 6 библиотек коротких прочетний, состоящие из одиночных прочтений, секвенированных на платформе IonTorrent PGM. Параметры сырых библиотек и библиотек, прошедших разные стадии фильтрации, приведены в таблице 3. Библиотеки содержат суммарно 28481151 прочтение, состоящие в общей сложности из 4685983175 нуклеотидов. Сырые библиотеки содержат в среднем по 4,75 млн прочтений и 78,1 млн оснований. Наименьший размер имеет библиотека Alm_2 – 3,06 млн прочтений, наибольший – библиотека A_bow_3 – 6,9 млн прочтений. В ходе удаления прочтений по параметрам длин и качества было удалено в примерно 8,2% всех прочтений.

Было проведено удаление прочтений, представляющих собой потенциальные фрагменты рРНК, двумя способами (см. раздел 2.1.1 – «Фильтрация библиотек»). Программа RNA-QC-Chain удалила в среднем 13,9% прочтений. Средний размер библиотек после такой фильтрации составляет 3,58 млн. прочтений, суммарный размер – 21,4 млн., что составляет 75,3% от исходного размера библиотек. Путём картирования на последовательности некодирующих РНК ячменя было удалено в среднем 8,05% прочтений. Средний размер библиотек после фильтрации с помощью картирования составляет 3,81

миллионов прочтений, суммарный размер очищенных библиотек – 22,9 миллионов прочтений, что составляет 80,4% от исходного размера библиотек.

Табл. 3. Характеристики библиотек коротких прочетний, использованных для анализа транскриптома i:Bw*Alm* и Bowman

Линия	Библиотека	Исходные библиотеки			Библиотеки после фильтрации	
		Кол-во прочтений	Общая длина, нуклеотидов	Средняя длина прочтения	Кол-во прочтений	Средняя длина прочтений
I:Bw*Alm*	Alm_1	4596395	720917644	172,1	3874912	166,94
	Alm_2	3056413	564667218	208,63	2372255	199,52
	Alm_3	5794644	992122000	181,91	5332600	181,47
Bowman	A_bow_1	4122599	689024046	177,37	2450068	175,49
	A_bow_2	4023501	350635368	128,56	2356572	126,56
	A_bow_3	6887599	1368616899	202,7	6523266	201,68

3.1.2. Картирование библиотек

Картирование производилось четырьмя способами, каждый из которых был применён к библиотекам, прошедшим три различных способа фильтрации. Основные результаты картирования библиотек приведены в таблице 4.

Как можно видеть из таблицы 4, четыре использованных подхода для картирования прочтений показывают разные результаты. Прежде всего, Dart успешно картирует в среднем 98,7% всех прочтений, в то время как STAR картирует в среднем 75,1% прочтений, HISAT2 – около 61,4% всех прочтений, а TopHat2 – около 33% всех прочтений. Dart уникально картирует в среднем 83,9% всех прочтений, STAR – 65,7%, Hisat2 – 55,3%, и TopHat2 уникально картирует 27,5% всех прочтений библиотек.

Таблица 4. Характеристики картирования библиотек коротких прочтений, секвенированных в ходе эксперимента с линией i:Bw*Alm*, на последовательность генома ячменя

Программа для картирования	Метод фильтрации библиотек	Картировано прочтений, %	Картировано уникально, %
Star	Prinseq	72,07	62,06
	RNA-QC-Chain	78,89	69,73
	Bowtie2	74,40	65,24
hisat2	Prinseq	59,80	52,49
	RNA-QC-Chain	64,55	58,76
	Bowtie2	59,75	54,70
Dart	Fil	98,70	79,05
	RNA-QC-Chain	98,64	87,40
	Bowtie2	98,73	85,32
TopHat	Prinseq	34,22	25,61
	RNA-QC-Chain	33,19	29,20
	Bowtie2	31,68	27,82

3.1.3 Поиск дифференциальной экспрессии генов

Перед идентификацией дифференциальной экспрессии генов было проведено удаление генов, имеющих экспрессию ниже пороговой. В результате для двенадцати имеющихся картирований библиотек было удалено от 20367 до 23450 генов. Поиск дифференциальной экспрессии был проведён тремя способами для каждого из 12 имеющихся картирований библиотек. Таким образом, было получено 36 наборов ДЭГ. Верификация дифференциальной экспрессии с помощью количественной полимеразной цепной реакции в реальном времени была проведена для пяти генов. Для этих генов была подсчитана корреляция значений изменения уровней экспрессии, определёнными из анализа библиотек, со значениями, полученными с помощью количественной ПЦР.

Было проведено ранжирование конвейеров (см. раздел 2.2.4 – «резюме биоинформатической обработки»), и были выбраны наиболее оптимальные конвейеры для обработки данных библиотек коротких прочтений и получения списка ДЭГ. Пять лучших

конвейеров приведены в таблице 5. Полный список конвейеров с соответствующими значениями приведён в дополнительной таблице 3.

Табл. 5. Конвейеры биоинформатической обработки, получившие наиболее высокую сумму рангов в эксперименте с линией i:Bw*Alm*

Метод фильтрации	Метод картирования	Метод поиска ДЭГ	Приоритет
bowtie2	Dart	EdgeR	84
bowtie2	Dart	DEGseq	84
Prinseq	Dart	DEGseq	83
bowtie2	Dart	DESeq2	81
Prinseq	Dart	DESeq2	76

Как можно видеть из таблицы 5, два конвейера получили наивысший приритет из всех. Первый конвейер состоит из следующих стадий: фильтрация библиотек по длинам и качеству прочтений с помощью программы Prinseq, удаление из библиотек последовательностей рибосомной РНК путём картирования библиотек программой Bowtie2 на референс в виде последовательностей рРНК, картирования полученных очищенных библиотек программой Dart и поиска дифференциальной экспрессии пакетом edgeR. Второй конвейер состоит из стадий: фильтрация библиотек по длинам и качеству прочтений с помощью программы Prinseq, удаление из библиотек последовательностей рибосомной РНК путём картирования библиотек программой Bowtie2 на референс в виде последовательностей рРНК, картирования полученных очищенных библиотек программой Dart и поиска дифференциальной экспрессии пакетом DEGseq. Значения изменения уровней экспрессии генов, определённые первым из этих конвейеров, имеют более высокое значение коэффициента корреляции Пирсона с результатами, полученными с помощью количественной ПЦР. Таким образом, этому конвейеру отдаётся предпочтение, и он будет в дальнейшем использован для анализа дифференциальной экспрессии генов.

Метод главных компонент – один из способов уменьшения размерности данных, с помощью которого можно выделить основные факторы, вносящие вклад в наблюдаемую изменчивость. Кроме того, этот метод позволяет исследовать сходство или близость между

различными объектами в имеющихся многомерных данных. Анализ такого рода, проведённый для показателей качества конвейеров, позволил разделить все конвейеры по осям главных компонент. График, полученный в результате разделения конвейеров по первым двум главным компонентам показан на рисунке 11.

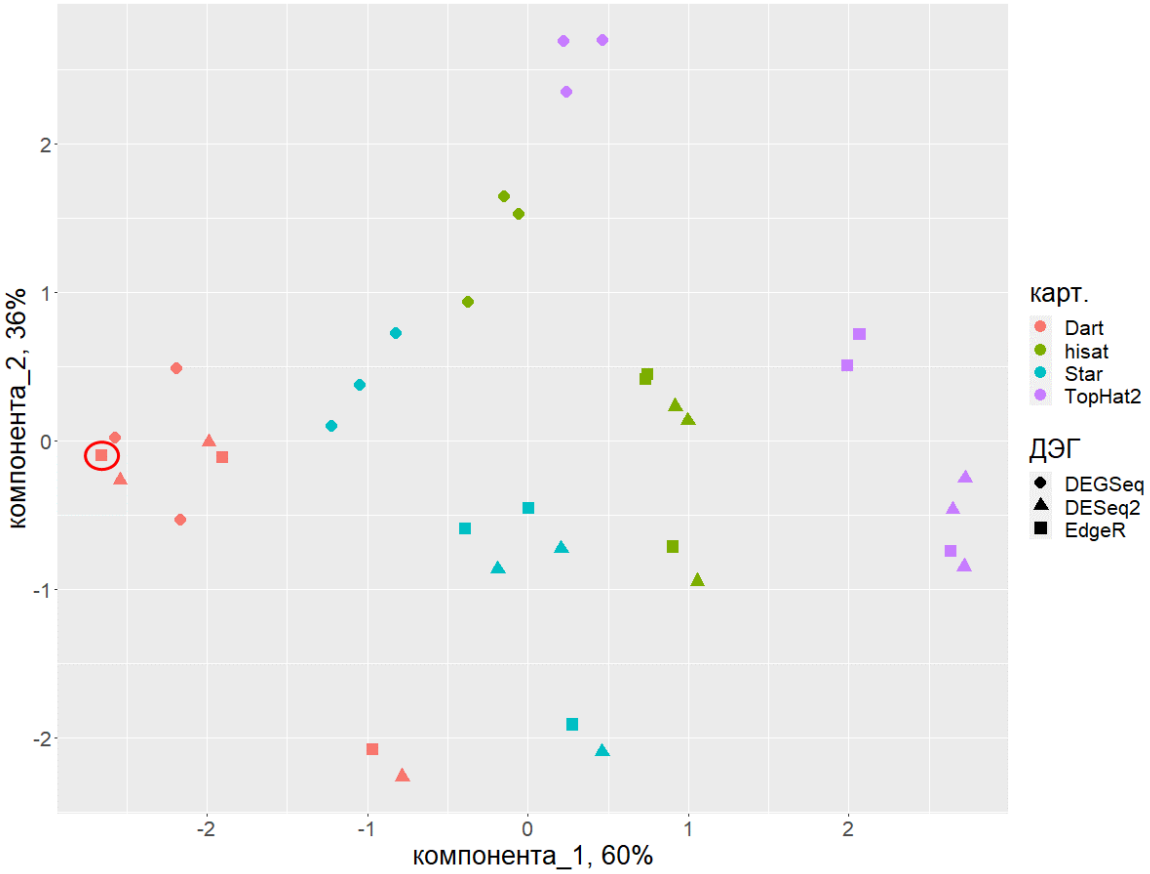

Рис. 11. Распределение конвейеров биоинформатической обработки библиотек коротких прочтений в эксперименте с линией i:Bw*Alm* по первой и второй главной компонентам. Красным цветом обведён конвейер биоинформатической обработки BowTie2-Dart-EdgeR, использованный в дальнейшей работе.

Суммарный вклад первых двух компонент в общую дисперсию составляет 96%. В первой главной компоненте веса переменных ‘n_cor’, ‘n_map’ и ‘n_uniq’ примерно одинаковы и составляют -0,53, -0,55 и -0,53, соответственно. Вес переменной ‘n_sdev’ составляет -0,35. Вес переменной ‘n_sdev’ во второй компоненте наибольший и составляет 0,67. Таким образом, по первой главной компоненте происходит разделение переменных в

зависимости как от метода картирования, так и от метода анализа дифференциальной экспрессии генов, в то время как по второй компоненте конвейеры разделяются в основном в зависимости от метода анализа ДЭГ.

При помощи этого конвейера для 19075 (47,8%) генов из 39841 аннотированных в текущей версии генома ячменя была обнаружена экспрессия не ниже порогового значения. 1365 генов (7,1% от всех экспрессирующихся генов, 3,4% от всех аннотированных генов ячменя) были определены как достоверно имеющие дифференциальную экспрессию. 78 (5%) из этих генов имеют повышенную экспрессию в линии i:Bw*Alm*, 1287 – пониженную экспрессию в этой линии. Изменение экспрессии генов проиллюстрировано на рисунке 12.

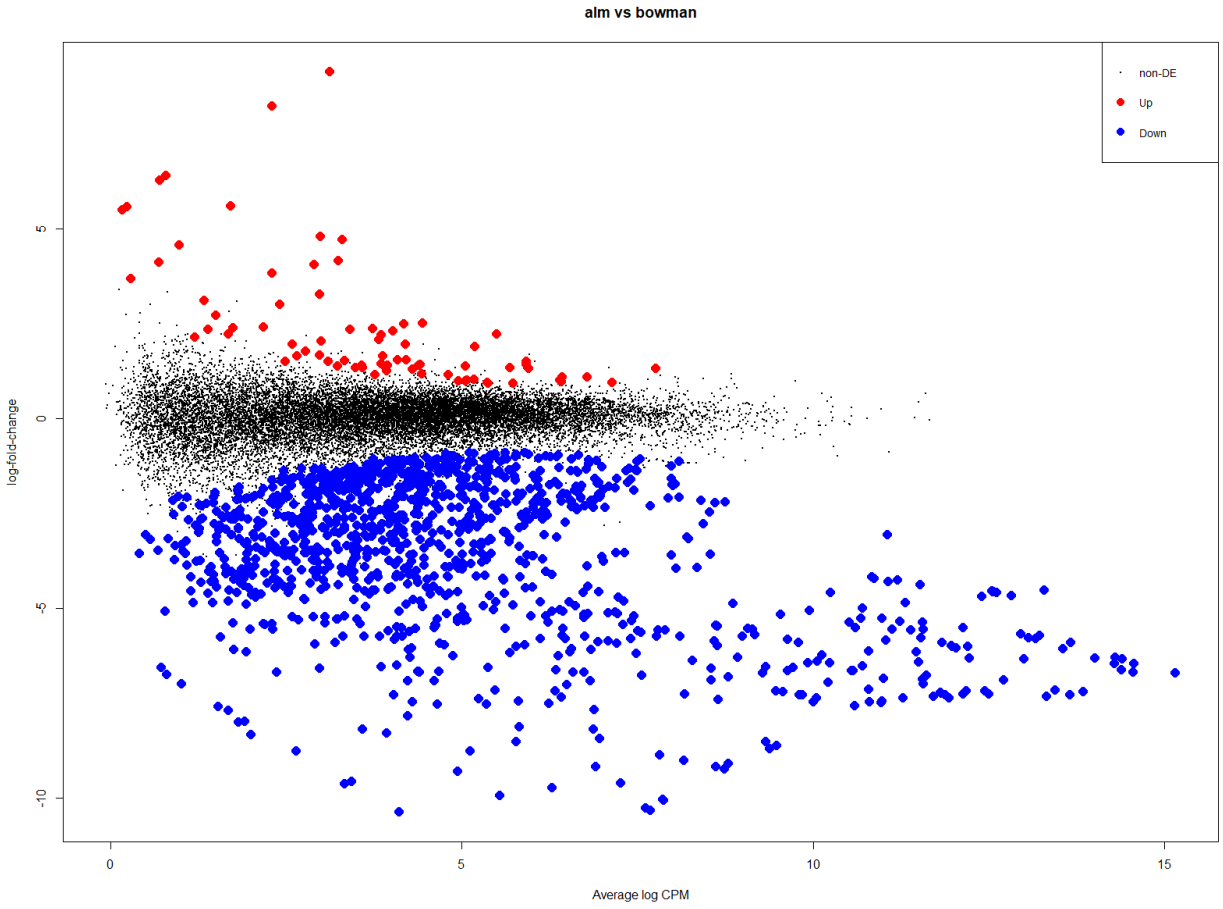

Рис. 12. Соотношение уровней экспрессии и изменения уровней экспрессии генов между линиями i:Bw*Alm* и Bowman

3.1.4 Анализ терминов генной онтологии

Из 1287 генов, имеющих пониженную экспрессию в линии i:Bw*Alm*, 796 встречаются в базе данных AgriGO. С помощью онлайн-сервиса SEA было выделено 234 терминов ГО как значимо обогащённые для этого списка генов. Из этих терминов 47, относятся к клеточной локализации белкового продукта гена, 54 относятся к молекулярной функции белкового продукта гена, и 133 терминов связаны с биологическими процессами.

Из числа генов с повышенной в линии i:Bw*Alm* экспрессией 52 встречаются в базе данных AgriGO. Было выделено 4 термина ГО, значимо обогащённых для этого списка, все 4 из которых относятся к категории «биологические процессы». Эти термины приведены в таблице 6.

Таблица 6. Термины генной онтологии, обогащённые для списка генов, имеющих повышенную экспрессию в линии i:Bw*Alm*.

Термин ГО	Описание	Кол-во ДЭГ	Кол-во генов	p-критерий	FDR
GO:0006952	Защитный ответ	5	143	$3{,}1 \cdot 10^{-5}$	0,00087
GO:0006950	Ответ на стресс	8	694	0,00029	0,004
GO:0050896	Ответ на стимул	10	1171	0,00051	0,0048
GO:0006508	Протеолиз	6	638	0,0048	0,033

3.1.5 Анализ метаболических путей, содержащих гены с дифференциальной экспрессией

В базе данных BarleyCyc представлено в общей сложности 496 метаболических путей, содержащих суммарно 3955 генов. Среди генов с пониженной в линии i:Bw*Alm* экспрессией 238 генов входят в общей сложности в 119 путей. При этом только 11 путей оказались статистически значимо представленными для этого списка генов. Эти пути перечислены в таблице 7. Наиболее достоверное обогащение ДЭГ наблюдается для метаболических путей «Цикл Кальвина-Бенсона» (CALVIN-PWY, $p = 1{,}68 \cdot 10^{-25}$), «Световая фаза фотосинтеза» (PWY-101, $p = 2{,}60 \cdot 10^{-24}$) и «Шунт РБФК» (PWY-5723, $p = 4{,}44 \cdot 10^{-16}$).

Таблица 7. Метаболические пути, значимо обогащённые для списка генов, имеющих пониженную в лемме ячменя линии i:Bw*Alm* по сравнению с леммой ячменя сорта Bowman экспрессию.

Метаболический путь	ID	Кол-во генов	Кол-во ДЭГ	Доля ДЭГ	FDR
Цикл Кальвина-Бенсона	CALVIN-PWY	61	35	57,4%	$1{,}68 \cdot 10^{-25}$
Световая фаза фотосинтеза	PWY-101	98	42	42,8%	$2{,}60 \cdot 10^{-24}$
Шунт РБФК	PWY-5723	63	28	44,4%	$4{,}44 \cdot 10^{-16}$
Аэробное дыхание I (Цитохром с)	PWY-3781	242	49	20,2%	$1{,}10 \cdot 10^{-12}$
Аэробное дыхание III (альтернативный оксидативный путь)	PWY-4302	184	38	20,6%	$1{,}12 \cdot 10^{-9}$
Фосфорилирование и дефосфорилирование NAD/NADH	PWY-5083	179	38	21,2%	$4{,}58 \cdot 10^{-10}$
de novo биосинтез аденозиновых нуклеотидов	PWY-7219	86	24	27,9%	$2{,}13 \cdot 10^{-8}$
Биосинтез L-глутамина	GLNSYN-PWY	12	7	58%	$5{,}34 \cdot 10^{-4}$
Биосинтез L-глутамина III	PWY-6549	44	12	27,3%	$1{,}9 \cdot 10^{-3}$
Цикл усвоения аммония II	PWY-6964	16	9	56,2%	$2{,}33 \cdot 10^{-5}$
Восстановление нитратов II (ассимиляционное)	PWY-381	14	7	50%	$2{,}05 \cdot 10^{-3}$

В базе данных BarleyCyc было обнаружено 15 метаболических путей, включающих в себя в общей сложности 12 генов с повышенной в линии i:Bw*Alm* экспрессией. Однако, ни один из этих путей не был статистически значимо представлен для этого списка генов.

3.1.6 Экспрессия генов пластома

Для трёх функциональных групп генов были оценены средние значения изменения уровней экспрессии и стандартные отклонения изменения уровней экспрессии. С помощью

теста Манна-Уитни была оценена достоверность различий между тремя функциональными группами. Значения изменения экспрессии, стандартных отклонений и достоверности различия между разными группами приведены на рисунке 13.

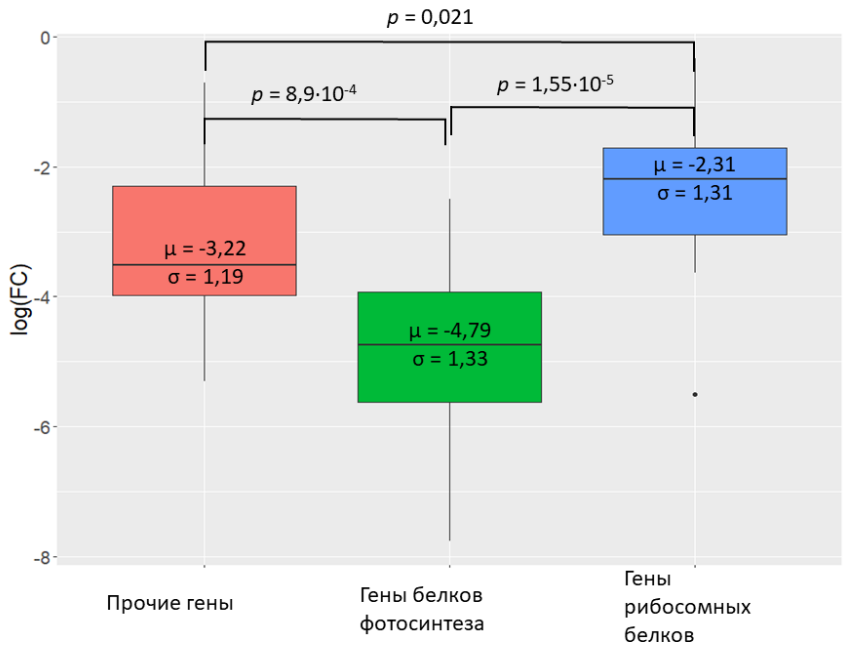

Рис. 13. Различия в изменениях экспрессии разных функциональных групп генов, локализованных в пластидном геноме ячменя

При сравнении групп «гены фотосинтеза» и «прочие гены» было получено значение $8{,}9 \cdot 10^{-4}$, при сравнении групп «рибосомные гены» и «прочие гены» было получено значение 0,021. Наконец, для групп «гены фотосинтеза» и «рибосомные гены» значение достоверности различий составляет $1{,}55 \cdot 10^{-5}$. Как можно видеть, различия в изменении уровней экспрессии между этими группами генов достоверны.

3.1.7 Экспрессия генов района Alm

Из числа этих генов 112 имеют экспрессию ниже установленного порога. Остальные 117 генов имеют экспрессию выше пороговой, что позволяет проводить сравнение уровней экспрессии. Из этих генов у семи была обнаружена достоверная дифференциальная экспрессия, причём шесть из этих генов понижают свою экспрессию в линии i:Bw*Alm*, и

один ген повышает экспрессию в этой линии. ДЭГ, локализованные в районе Alm, приведены в таблице 8.

Табл. 8. Гены, имеющие достоверную дифференцильную экспрессию между линиями ячменя i:Bw*Alm* и Bowman, локализованные в районе Alm. ИЭ – изменение экспрессии, значение логарифмировано по основанию 2

Ensembl ID гена	Описание белкового продукта	ИЭ	p-критерий
HORVU3Hr1G034230	40S рибосомный белок	2,32	$5{,}53 \cdot 10^{-4}$
HORVU3Hr1G034100	Rmlc-подобный белок суперсемейства купин-доменных белков	-0,92	$2{,}38 \cdot 10^{-2}$
HORVU3Hr1G032490	Глюкоза-6-фосфат транслокатор-подобный белок	-1,91	$7{,}25 \cdot 10^{-10}$
HORVU3Hr1G034640	слабая гомология с SANT/Myb домен-содержащим белком *Cinnamomum micranthum* f. *Kanehirae*	-1,93	$2{,}02 \cdot 10^{-3}$
HORVU3Hr1G034070	слабая гомология с гистоном 2A *Erysiphe pulchra*	-5,65	$1{,}56 \cdot 10^{-3}$
HORVU3Hr1G034060	Гистон H2B	-6,08	$1{,}28 \cdot 10^{-3}$
HORVU3Hr1G031940	Субъединица 2 протеазы внутренней мембраны митохондрий	-8,63	$1{,}27 \cdot 10^{-24}$

Ген 40S рибосомного белка (EnsemblID HORVU3Hr1G034230) обращает на себя внимание тем, что это единственный из локализованных в районе Alm генов, достоверно повышающий свою экспрессию в линии i:Bw*Alm* по сравнению с линией Bowman. Экспрессия этого гена в 4,99 ($\log_2$(FC) = 2,32) раз выше в растениях линии i:Bw*Alm*. Гены, понижающие экспрессию в линии i:Bw*Alm*, кодируют следующие белки: белок суперсемейства купин-доменных белков, глюкоза-6-фосфат транслокатор-подобный белок, гистон H2B, гомолог гистона 2A, транскрипционный фактор из семейства SANT/Myb и субъединицу 2 протеазы, локализованной на внутренней мембране митохондрий.

3.1.8 Реконструкция транскриптома *de novo*

Для линии ячменя i:Bw*Alm* и использованного в качестве контроля сорта Bowman были получены четыре индивидуальные сборки *de novo* транскриптома леммы и перикарпа и одна мета-сборка, составленная из четырёх индивидуальных сборок. В таблице 9 приведены результаты сборки *de novo* транскриптома ячменя линий i:Bw*Alm* и Bowman, включая мета-сборки, а также общей для двух линий генеральной сборки.

Мета-сборка транскриптома ячменя линий i:Bw*Alm* и Bowman, полученная из сборок *de novo*, созданных с помощью RNA-Spades, Trans-ABySS и Trinity и геном-ориентированной сборки Trinity, до удаления избыточности состоит из 169232 контигов. Не-избыточная мета-сборка состоит из 68414 контигов суммарной длиной 46440750 оснований. Наибольшая длина контига в сборке – 9920 нуклеотидов, средняя длина – 678,8 нуклеотидов, N50 – 936 нуклеотидов. Удаление избыточности уменьшило размер мета-сборки до 40,4% от исходного.

Табл. 9. Характеристики *de novo* сборок транскриптома леммы ячменя линий Bowman и i:Bw*Alm*

Сборка	Размер сборки, контигов		N50 длин контигов	Средняя длина контига	Прочтений картировано, %
	Избыточная	Не-избыточная			
Abyss	705015	40806	1076	723,6	67,08
Spades	22649	19181	1130	1072,65	39,13
Trinity	267201	52005	976	741,19	64,97
GG	451309	57240	766	594,82	61,37
Мета-сборка	169232	68414	936	678,82	61,47

Наибольшая доля прочтений из библиотек была выравнена на сборку транскриптома abyss, в то время как наименьшее – на сборку spades. На мета-сборку транскриптома было выровнено 61,47% всех коротких прочтений библиотек RNA-seq. Кроме того, с помощью программы TransRate известные CDS ячменя были обнаружены в сборках транскриптома *de novo*. Результаты идентификации CDS приведены для разных сборок в таблице 10.

Наибольшее количество известных CDS – 29790 – было обнаружено в мета-сборке транскриптома. Также, в мета-сборке было обнаружено самое большое количество CDS, покрытых контигами сборки не менее чем на 95%. Однако, при этом наибольшая доля контигов, для которых была обнаружена значимая гомология с CDS ячменя, представлена в сборке spades – 90,3%. В мета-сборке этот показатель составляет всего 62,7%, что меньше чем во всех индивидуальных сборках.

Табл. 10. Количество известных CDS ячменя, обнаруженных в *de novo* сборках эксперимента alm. Указано количество и доля контигов, для которых была обнаружена гомология с известными CDS ячменя, количество CDS, имеющих гомологию к контигам сборки транскриптома, и количество контигов, не менее 95% длины которых гомологичных известным CDS.

Сборка	Контигов, гомологичных CDS	% контигов, гомологичных CDS	CDS найдено	p_95
Abyss	30530	0,748	22420	2542
Spades	17323	0,903	14989	644
Trinity	35547	0,684	27173	1779
GG	38686	0,676	26978	2240
Мета-сборка	42887	0,627	29790	3073

Далее, для оценки вклада каждого из сборщиков в структуру мета-сборки транскриптома была проведена оценка перекрывания множеств CDS ячменя, встреченных в каждой из индивидуальных сборок. Полученные результаты проиллюстрированы на рисунке 14.

Как можно видеть, 7191 CDS ячменя были обнаружены во всех четырёх индивидуальных сборках транскриптома, ещё 9305 CDS были обнаружены в трёх сборках из четырёх. 14615 CDS были обнаружены только в одной из четырёх сборок, из которых наибольшее количество – 5137 – были обнаружены только в сборке Trinity, наименьшее – 2086 – только в сборке Spades. Наибольшее перекрывание множеств обнаруженных CDS наблюдается между индивидуальными сборками Trinity и GG – 18258 CDS.

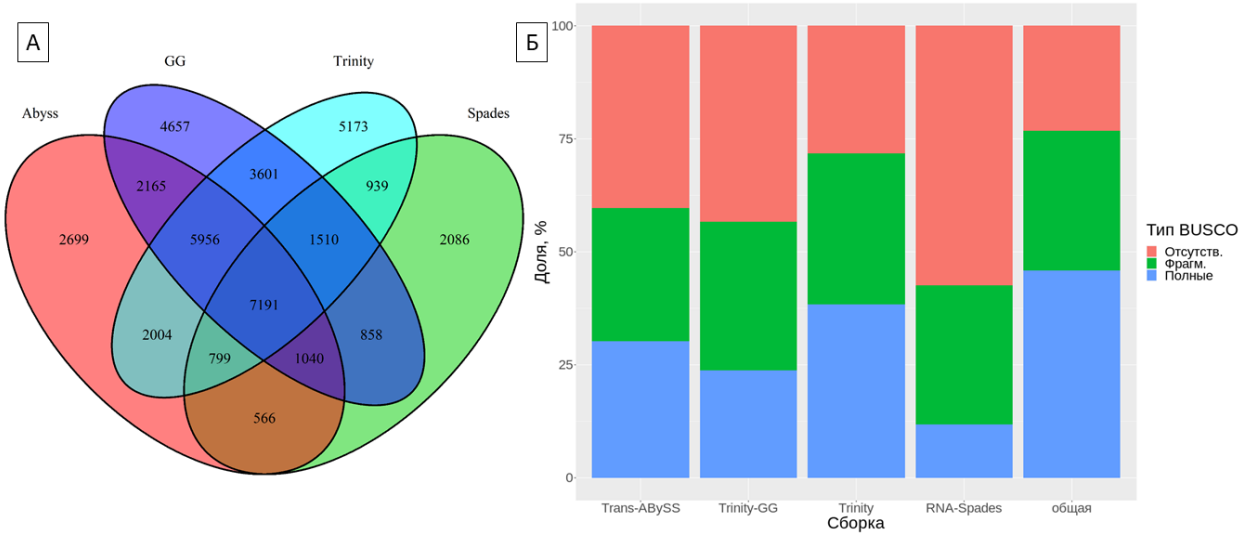

Рисунок 14. (А) Диаграмма Венна, показывающая перекрывание множеств CDS, обнаруженных в индивидуальных сборках транскриптома *de novo* в эксперименте alm. (Б) Значения критерия BUSCO для разных сборок транскриптома в эксперименте Alm

В контигах каждой из сборок были идентифицированы открытые рамки считывания. Обнаруженные в контигах общей сборки ОРС кодируют 58636 белковых продуктов длинами не менее 30 аминокислотных остатков. Критерий полноты BUSCO для мета-сборки составляет 78,6%, что больше, чем для любой из индвидуальных сборок транскриптома. При этом, доля обнаруженных полных последовательностей – 45,9%, а доля фрагментированных – 30,9%.

Оценка уровней экспрессии показала, что 33257 контигов имеют уровни экспрессии выше порогового, и для этих контигов можно провести оценку изменения экспрессии. Было выявлено 652 контига, понижающих экспрессию, и 110 контигов, повышающих экспрессию в линии i:Bw*Alm*.

Установлено, что 943 контига из мета-сборки транскриптома не имеют значимой гомологии к последовательности генома ячменя. Каждый из этих контигов имеет по одной открытой рамке считывания. Из этих контигов 556 были удалены как артефакты сборки и 235 – как потенциальная контаминация чужеродным генетическим материалом. Из 152 новых контигов, взятых в дальнейшую обработку, 5 имеют достоверную дифференциальную экспрессию. Из них 3 повышают уровень экспрессии в линии i:Bw*Alm*,

два других – понижают. Эти контиги, их изменение экспрессии и обнаруженные для них лучшие гомологи приведены в таблице 11.

Онлайн-сервис поиска доменной структуры NCBI Structure не обнаружил в аминокислотных продуктах контигов DN20386 и DN16917 консервативных доменов. Кодирующий потенциал этих контигов, подсчитанный с помощью онлайн-сервиса cpc2, составляет 0,17 и 0,41, соответсвенно. Онлайн-сервис InterProScan определил наличие в этих аминокислотных продуктах структуры ‘Mobi_db’ [Necci и др., 2017], связанной с так называемыми внутренне неупорядоченными белками. В аминокислотном продукте контига GG_19862 был обнаружен каталитический домен протеинкиназы ($E = 2{,}68 \cdot 10^{-10}$). Его кодирующий потенциал оценен как 0,16. InterProScan также обнаружил в его структуре домен белков из суперсемейства протеинкиназ (IPR011009). В аминокислотном продукте контига DN4161 обнаружен с помощью сервиса NCBI Structure РНК-связывающий домен второго типа I K-гомологов ($E = 1{,}3 \cdot 10^{-5}$). Однако, сервис InterProScan обнаружил в нём только структуру ‘Mobi_db’. Его кодирующий потенциал оценён как 0,99.

Наконец, в аминокислотном продукте контига DN2647 было обнаружено наиболее достоверное наличие доменной структуры – домен белкового семейства прохибитинов, суперсемейства SPFH ($E = 1{,}49 \cdot 10^{-13}$). Доменная структура аминокислотного продукта представлена на рисунке 14. Для него также наблюдается наиболее достоверная из всех гомология с функционально аннотированным белком – dao-5-подобным ядрышковым белком *T. dicoccoides* ($E = 1{,}32 \cdot 10^{-69}$). Помимо этого, он имеет гомологию к прохибитин-1-подобному белку *Solanum pennellii* со значением ($E = 3 \cdot 10^{-20}$). Это согласуется с результатами InterProScan, который обнаружил в этом белковом продукте доменную структуру Band_7 (IPR001107) и отнёс его к семейству прохибитинов (IPR000163) и суперсемейству Band_7/SPFH_dom_sf (IPR036013), включающему в себя прохибитины. Кодирующий потенциал этого белка оценён как 0,23. Отметим, однако, что онлайн-сервис cpc2 определил размеры белкового продукта этого транскрипта как 196 аминокислот, в то время как определённый с помощью EvidentialGene размер белкового продукта составляет 230 аминокислот.

Табл. 11. Контиги в сборке транскриптома линии i:Bw*Alm*, не имеющие значимой гомологии к последовательности генома или транскриптома ячменя.

Контиг	Экспрессия		Лучший гомолог из базы данных NCBI Protein		
	ДЭ	FDR	Идентификатор	Описание	E-критерий
DN20386	-2,48	0,019	XP_022087303.1	Тубулин полиглутамилаза TTLL7-подобная, *Acanthaster planci*	7,98
GG_19862	-3,32	0,018	XP_020190453.1	Серин/треонин киназа белков RHS3, *A. tauschii*	$2{,}29 \cdot 10^{-6}$
DN16917	2,3	0,043	WP_239129868.1	GNAT семейства N-ацетилтрансфераза, *Planobispora takensis*	4,38
DN2647	5,67	0,0028	XP_037411501.1	Ядрышковый белок dao-5-подобный, *Triticum dicoccoides*	$1{,}32 \cdot 10^{-69}$
DN4161	7,09	$1{,}3 \cdot 10^{-5}$	XP_015623976.1	РНК-связывающий KH домен-содержащий белок PEPPER, *O. sativa*	$1{,}75 \cdot 10^{-18}$

Для этого транскрипта была проведена локализация в геноме ячменя путём сравнения изогенной линии i:Bw*Alm*, сорта Bowman и пшенично-ячменных замещённых линий. При использовании пары праймеров, специфичных для данного транскрипта, амплификация происходит на генетическом материале ДНК ячменя линии i:Bw*Alm*, тогда как у сорта Bowman продукт амплификации отсутствует. При сравнении пшенично-ячменных линий специфический фрагмент наблюдался у линии, дополненной хромосомой 3H, и у дителосомной линии 3HS (Рис. 15, A).

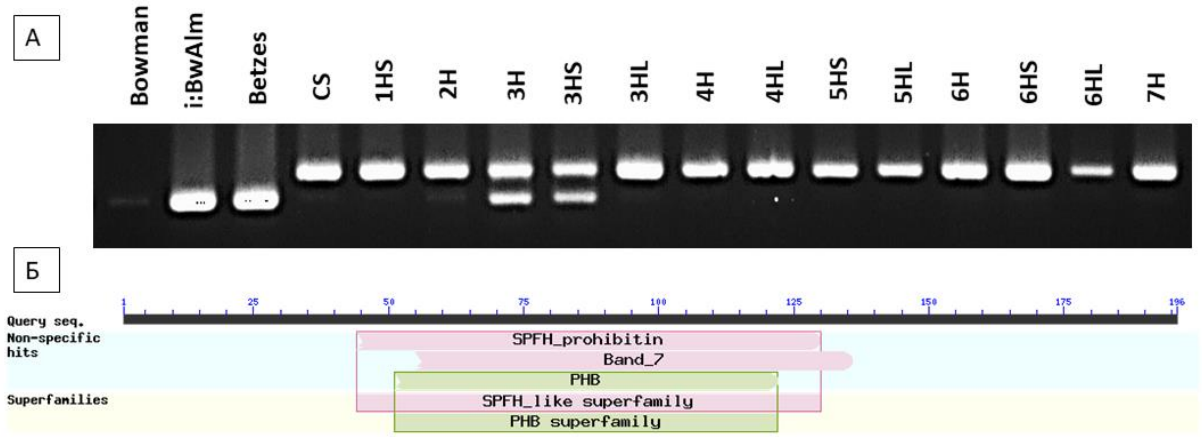

Рисунок 15. (А) – продукты амплификации, полученные при использовании пары праймеров, специфичной для контига DN2647, на генетическом материале пшенично-ячменной линии (Б) – Доменная структура обнаруженного в линии i:Bw*Alm* транскрипта DN2647

3.2 Анализ транскриптома почти изогенной линии ячменя i:Bw*Blp* в сравнении с Bowman

Был проведён биоинформатический анализ библиотек коротких прочтений, полученных в эксперименте с линией ячменя i:Bw*Blp*. Почти изогенная линия i:Bw*Blp*, полученная на основе сорта Bowman, контрастна с ним по окраске колоса – плотно прилегающая к перикарпу зерна цветковая чешуя (лемма) этой линии содержат меланин, что придаёт колосу чёрный цвет. Меланин встречается в цветковой чешуе дикорастущих растений, и защищает их от механических и температурных повреждений и поражения патогенами и паразитами. Формирование этого фенотипа обусловлено действием гена Blp, локализованного в длинном плече хромосомы 1H. На данный момент последовательность и функции этого гена неизвестны. Сравнительный транскриптомный анализ леммы ячменя линии i:Bw*Blp* и сорта Bowman, следовательно, может помочь в определении структуры гена *Blp* и помочь в понимании синтеза меланина у злаков и генетического контроля этого процесса.

3.2.1 Предобработка библиотек коротких прочтений

Шесть библиотек коротких прочтений, секвенированных на платформе IonTorrent Ion Proton, были взяты для данного анализа. Библиотеки содержат в общей сложности

23128312 прочтений и 4034062730 нуклеотидов. Наибольший размер имеет библиотека ‘B_bow_3’ – 5253524 прочтений; наименьший размер у библиотеки ‘B_bow_1’ – 1769261 прочтений. Средний размер библиотек – 3,85 млн прочтений, средняя длина прочтений – 174,4 нуклеотида. В процессе фильтрации было удалено в общей сложности около 12,5% всех прочтений. После фильтрации библиотеки имеют средний размер 3,38 млн. прочтений. Общие характеристики использованных библиотек можно найти в таблице 12.

Табл. 12. Характеристики библиотек коротких прочетний, использованных для анализа транскриптома почти изогенной линии ячменя i:Bw*Blp*

Линия	Библиотека	Исходные библиотеки			Библиотеки после фильтрации	
		Размер, прочтений	Размер, нуклеотидов	Средняя длина	Размер, прочтений	Средняя длина
BLP	Blp_1	3583148	638776453	178,27	1311442	185,39
	Blp_2	4710862	741481190	157,4	1687289	156,96
	Blp_3	4070591	564284897	138,62	1864073	146,02
Bowman	B_bow_1	1769261	241954662	136,75	438702	164,66
	B_bow_2	3740926	762252900	203,76	1092191	199,48
	B_bow_3	5253524	1085312628	206,59	2364034	209,00

В каждой из библиотек была обнаружена последовательность, встречающаяся не менее чем в 1% всех прочтений, входящих в эту библиотеку. Наиболее перепредставленной оказалась последовательность ‘TACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCC’, встречающаяся в 2,9% всех прочтений библиотеки BLP2, и не менее чем в 1% всех прочтений в каждой другой библиотеке. С помощью онлайн-сервиса blastn эта последовательность была выровнена на нуклеотидные последовательности, содержащиеся в базе данных NCBI Nucleotide, и была обнаружена полная идентичность с последовательностью внутреннего транскрибируемого спейсера гена 18S рибосомной РНК неопределённого вида из рода *Debaryomyces*. Последовательность ‘GCGACCCCAGGTCAGGCGGGACTACCCGCTGAGTTTAAGCATATAAATAA’ встречается в 2,3% всех прочтений в библиотеке Blp_2, и 1,06% всех прочтений в

библиотеке B_bow_3, и является наиболее сильно перепредставленной последовательностью во всех четырёх остальных библиотеках. Эта последовательность имеет полную идентичность последовательностям 5,8S рибосомных РНК, представленных в базе данных NCBI nucleotide, относящихся к разным видам, в том числе нескольким видам злаков: *Triticum urartu*, *Triticum tugridum*, *Triticum monococcum*, *Triticum aestivum*, *Aegilops cylindrica*, *Aegilops crassa*. Всё это говорит о содержании в библиотеках рибосомной РНК и необходимости их дополнительной очистки. Такая очистка была проведена с помощью программы RNA-QC-Chain и путём картирования программой Bowtie на последовательности некодирующих РНК (см. раздел «методы»).

После удаления рРНК с помощью программы RNA-QC-Chain размер библиотек уменьшился в среднем до 43% от размера чистых библиотек. В результате удаления рРНК путём картирования на последовательности некодирующих РНК было удалено в среднем 36,4% всех прочтений, составлявших чистые библиотеки.

3.2.2 Картирование библиотек

Картирование было проведено четыремя программами – Dart, Star, Hisat2 и TopHat2. Картирование было проведено каждой из программ отдельно каждого из наборов очищенных библиотек – не прошедших удаление рРНК, прошедших удаление рРНК с помощью RNA-QC-Chain, и очищенных от рРНК картированием на последовательности рибосомных РНК ячменя. Итого было получено 12 картирований для каждой из 6 библиотек. Основные характеристики картирований приведены в таблице 13.

Из таблицы 13 видно, что производительность использованных для картирования средств отличается. TopHat2 является как наименее чувствительной, так и наименее точной программой для картирования коротких прочтений на референс. Программа Star имеет более высокий уровень чувствительности – картирует 85% прочтений из библиотек, не прошедших удаление рРНК, 84% прочтений из библиотек, прошедших удаление рРНК картированием и 90% прочтений из библиотек, обработанных RNA-QC-Chain. Dart, в то же время, картирует более 98% всех прочтений библиотек, независимо от способа предобработки.

Таблица 13. Характеристики картирования библиотек коротких прочтений в эксперименте с линией i:Bw*Blp*

Программа для картирования	Метод фильтрации библиотек	Картировано прочтений, %	Картировано уникально, %
Star	Prinseq	85,09	34,67
	RNA-QC-Chain	90,24	74,39
	Bowtie2	84,21	65,61
hisat2	Prinseq	54,78	34,42
	RNA-QC-Chain	77,64	68,39
	Bowtie2	67,23	59,95
Dart	Prinseq	99,12	40,50
	RNA-QC-Chain	98,28	82,19
	Bowtie2	98,27	74,50
TopHat2	Prinseq	50,60	15,89
	RNA-QC-Chain	44,37	36,85
	Bowtie2	35,45	24,40

3.2.3 Поиск дифференциальной экспрессии генов

Количество прочтений, картированных на каждый из генов, было подсчитано с помощью программы featureCounts из пакета программ Subread. Далее, гены, имеющие низкую экспрессию, были удалены из рассмотрения, согласно процедуре, описанной в разделе «Методы».

Три пакета для языка R были использованы для определения генов, имеющих статистически значимую дифференциальную экспрессию: EdgeR, DEGSeq и DESeq2. Таким образом, в общей сложности было опробовано 36 конвейеров биоинформатической обработки данных эксперимента RNA-seq. Для ряда генов была проведена верификация дифференциальной экспрессии с помощью полимеразной цепной реакции в реальном времени.

Наибольшее значение коэффициента корреляции Пирсона между идентифицированными в результате биоинформатического анализа RNA-seq и наблюдаемыми с помощью количественной ПЦР в реальном времени значениями

изменения экспрессии генов получено для конвейеров обработки, включающих в себя картирование с помощью программы Hisat и поиск дифференциальной экспрессии с помощью пакета EdgeR. При этом, способ фильтрации библиотек существенного влияния на полученные результаты не оказывает.

Была проведена приоретизация конвейеров биоинформатической обработки (см. раздел 2.2.4 – «резюме биоинформатической обработки»). Полные список конвейеров с соответствующими значениями приведён в дополнительной таблице 4. В таблице 14 показаны конвейеры, получившие наивысший приоритет.

Таблица 14. Приоретизация конвейеров биоинформатической обработки в эксперименте с линией i:Bw*blp*

Метод фильтрации	Метод картирования	Метод поиска ДЭГ	Сумма рангов
Bowtie	Hisat2	edgeR	84
Chain	Dart	edgeR	83
Chain	Dart	deseq2	77
Prinseq	Dart	edgeR	77
Bowtie	Dart	deseq2	76

Наибольший приоритет получил конвейер биоинформатической обработки, состоящий из следующих стадий: удаление рибосомных РНК из библиотек коротких прочтений путём картирования на референс из последовательностей рибосомных РНК ячменя; картирование очищенных библиотек на геном ячменя с помощью программы Hisat2; поиск дифференциальной экспрессии с помощью пакета EdgeR для языка R. Использованный конвейер обработки представлен на рисунке 16.

В первой главной компоненте наибольшие веса имеют исходные переменные ‘n_tot’ и ‘n_sdev’: -0,68 и -0,59, соответственно. Веса двух других переменных составляют -0,27 и -0,34. Во второй компоненте наибольшие веса имеют исходные переменные ‘n_map’ и ‘n_uniq’: -0,64 и -0,6, соответственно; веса переменных ‘n_tot’ и ‘n_sdev’ составляют 0,2 и 0,42. Таким образом, по первой главной компоненте происходит разделение конвейеров в

основном в соответствие с качеством определения дифференциальной экспрессии, в то время как по второй компоненте – в соответствие с качеством картирования библиотек.

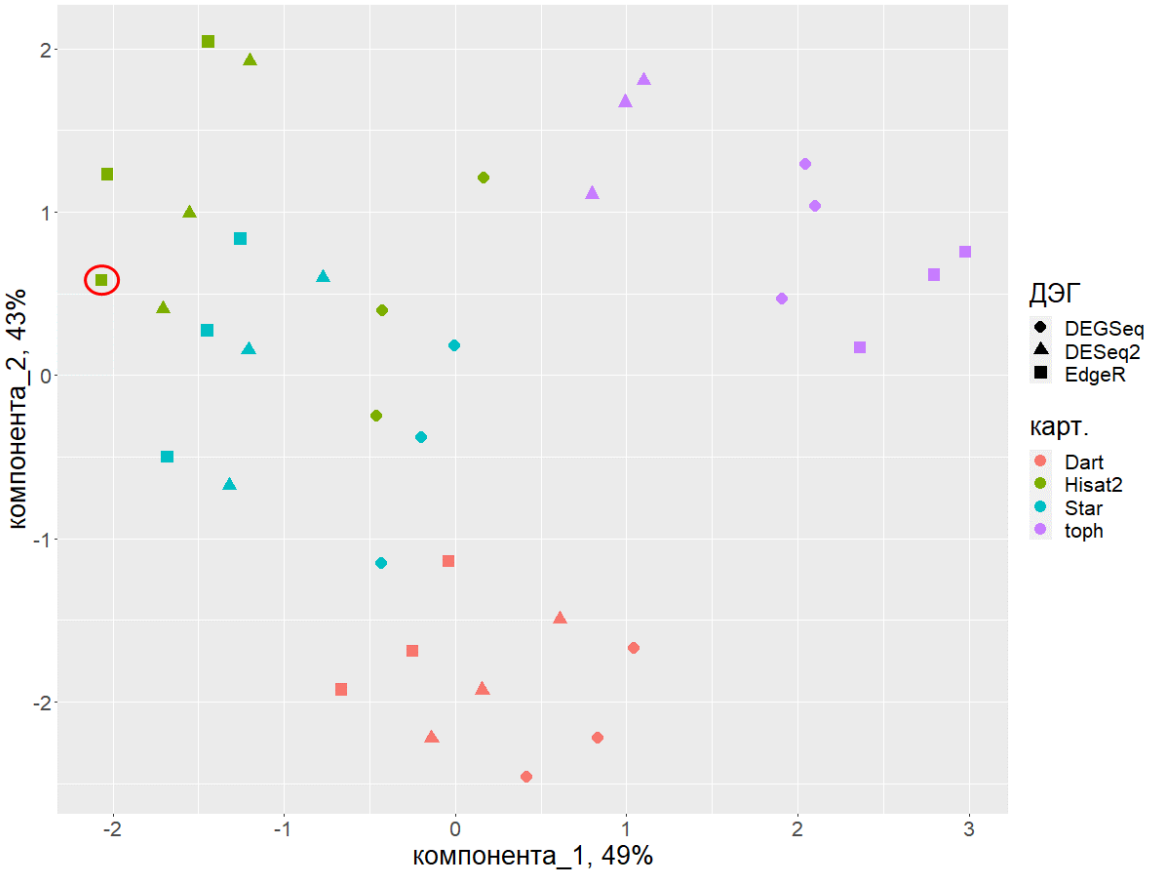

Рисунок 16. Распределение конвейеров биоинформатической обработки библиотек линии i:Bw*Blp* и Bowman по первым двум главным компонентам. Красным цветом выделен конвейер обработки Bowtie2-Hisat2-EdgeR, использованный в дальнейшей работе

Конвейер биоинформатической обработки данных, состоящий из следующих стадий: фильтрация библиотек с помощью программы Prinseq, удаление рРНК путём картирования библиотек на последовательности рРНК ячменя, картирования библиотек на геном с помощью программы Hisat2 и поиск дифференциальной экспрессии генов с помощью пакета EdgeR, имеет максимальную среди всех конвейеров сумму рангов. Этот конвейер был в дальнейшем использован для анализа дифференциальной экспрессии генов. С помощью этого конвейера было установлено, что 794 гена повышают экспрессию у линии i:Bw*blp* по сравнению с Bowman, в то время как 480 генов – понижают экспрессию в этой

линии. MA-график, иллюстрирующий соотношение генов с повышенной и пониженной в линии i:Bw*Blp* по сравнению с Bowman экспрессией, представлен на рисунке 17.

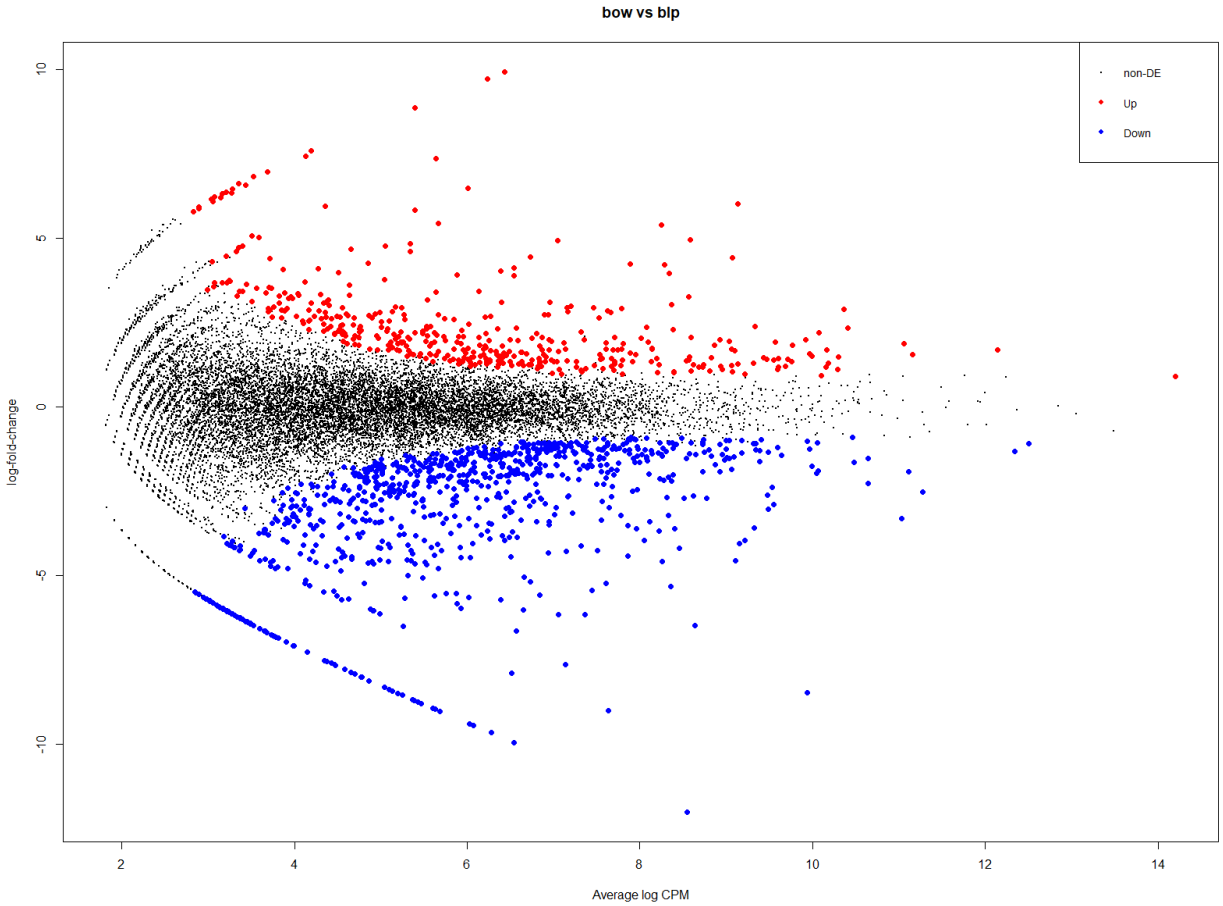

Рисунок 17. Соотношение общих уровней экспрессии и изменения экспрессии генов между i:Bw*Blp* и Bowman

3.2.4 Анализ терминов генной онтологии

Для 288 генов с пониженной в линии i:Bw*Blp* экспрессией и 565 генов с повышенной в этой линии экспрессией была обнаружена аннотация в базе данных AgriGO. С помощью анализа SEA было обнаружено 48 терминов генной онтологии, значимо представленных для генов с пониженной экспрессией и 67 терминов, обогащённых для генов с повышенной экспрессией у i:Bw*Blp*. При этом, 18 терминов ГО, обогащённых для списка генов с пониженной экспрессией у i:Bw*Blp*, относятся к категории «биологический процесс», 16 – к категории «молекулярная функция», 14 – к категории «клеточная локализация»; 39 терминов, обогащённых для набора генов с повышенной экспрессией у i:Bw*Blp*, относятся

к категории «биологический процесс» и 28 – к категории «молекулярная функция». Терминов генной онтологии, входящих в категорию «клеточная локализация», обогащённых для списка генов с повышенной экспрессией в линии i:Bw*Blp*, обнаружено не было.

Для генов с пониженной экспрессией у i:Bw*Blp* были достоверно обогащены ГО термины «Фотосистема» (p = 0,0076), «тилакоид» (p = 0,013), «фотосинтез» (p = 0,0098). Кроме того, для генов с пониженной экспрессией в этой линии значимо обогащены термины: «метаболизм клеточных полисахаридов» (p = 0,034), «метаболизм моносахаридов» (p = 0,0011), «связывание ионов магния» (p = 0,047). Наконец, для этого списка генов обогащены также термины «метаболический процесс альфа-аминокислот» (p = $9 \cdot 10^{-4}$), «ГТФазная активность» (0,0061), «микротрубочки» (p = $1{,}4 \cdot 10^{-4}$).

Для генов с повышенной экспрессией в линии i:Bw*Blp* обогащены термины генной онтологии, связанные с биосинтезом различных соединений: «биосинтез жирных кислот» (p = $5{,}910^{-4}$), «биосинтез изопреноидов» (p = $2 \cdot 10^{-4}$), «метаболизм триптофана» (p = $1{,}3 \cdot 10^{-3}$), «биосинтез альфа-аминокислот» (p= 0,002); термины ГО, относящиеся к связыванию различных субстратов – «связывание ионов железа» (p = $1{,}2 \cdot 10^{-4}$), «связывание ионов магния» (p = 0,016), «связывание флавинадениндинуклеотидов» (p = 0,016); термины ГО, относящиеся к окислительно-восстановительной активности, затрагивающей при этом ряд субстратов, в число которых входят дифенолы (p = 0,016) и CH-OH группы (p = 0,016). Также, обогащены термины ГО, связанные с ингибированием эндопептидазной активности: «активность ингибиторов эндопептидаз» (p = 0,016), «активность ингибиторов сериновых эндопептидаз» (p = 0,017).

3.2.5 Метаболические пути

В базе данных BarleyCyc был обнаружен 193 метаболических пути, включающий в себя в общей сложности 179 гена с повышенной в линии i:Bw*Blp* экспрессией. Однако, статистически значимо обогащённых путей обнаружено не было. Наибольшее значение достоверности наблюдается для метаболических путей «биосинтез *o*-дихинонов» (p = $1{,}59 \cdot 10^{-3}$), «первичные реакции биосинтеза фенилпропаноидов» (p = $3{,}36 \cdot 10^{-3}$), «биосинтез L-триптофана» (p = $4{,}58 \cdot 10^{-3}$). Однако, после поправки на множественное сравнение значения достоверности для обогащения этих метаболических путей ДЭГ превысили

установленный порог 0,05. В этих метаболических путях участвуют 3, 4 и 5 генов с пониженной у линии i:Bw*Blp* экспрессией, соответственно. Отметим, что в базе данных BarleyCyc аннотировано всего 6 генов, входящих в метаболический путь «биосинтез о-дихинонов», и, таким образом, половина из этих генов повышает экспрессию в линии i:Bw*Blp*.

Наибольшее количество ДЭГ с повышенной экспрессией у линии i:Bw*Blp* встречено в метаболических путях «биосинтез мономеров суберина» (7 ДЭГ), «биосинтез бензоата I» (6 ДЭГ), «биосинтез флавонола» (6 ДЭГ), «биосинтез флавоноидов» (6 ДЭГ). Достоверность обогащения этих метаболических путей также превышает установленный порог 0,05 после поправки на множественное сравнение. Кроме того, по 3 ДЭГ было обнаружено в метаболических путях «биосинтез кутина» и «биосинтез кутикулярных восков».

В то же время, было обнаружено 108 метаболических путей, включающих ДЭГ с пониженной экспрессией у линии i:Bw*Blp*, и 3 из этих метаболических путей были статистически значимо представлены. Они приведены в таблице 15.

Таб. 15. Метаболические пути, значимо обогащённые для генов с пониженной экспрессией у линии i:Bw*Blp*.

Метаболический путь	Кол-во ДЭГ	Кол-во генов	Доля ДЭГ	FDR
Шунт РБФК	10	63	15,8%	$2.43 \cdot 10^{-3}$
Цикл усвоения аммония II	6	16	37,5%	$9.56 \cdot 10^{-4}$
Цикл Кальвина-Бенсона	15	61	27,8%	$8.82 \cdot 10^{-11}$

Как можно видеть из таблицы 15, 15 генов, входящих в путь «цикл Кальвина-Бенсона» и 10 генов, входящих в метаболический путь «шунт РБФК», понижают свою экспрессию у линии i:Bw*Blp*. Значимость обогащения этих путей составляет $1{,}5 \cdot 10^{-8}$ и $2{,}51 \cdot 10^{-3}$, соответственно. Шесть генов из шестнадцати, входящих в метаболический путь

«цикл усвоения аммония II», понижают свою экспрессию у линии i:Bw*Blp* по сравнению с Bowman, значимость обогащения этого пути составляет $1{,}1 \cdot 10^{-3}$.

3.2.6 Экспрессия генов пластома

Анализ экспрессии генов пластид, относящихся к разным функциональным группам, показал различия в уровнях изменения экспрессии между этими группами. Средние значения изменения уровней экспрессии и среднеквадратичные отклонения для разных функциональных групп генов пластидного генома, а также различия в изменении экспрессии для разных функциональных групп генов пластома показаны на рисунке 18.

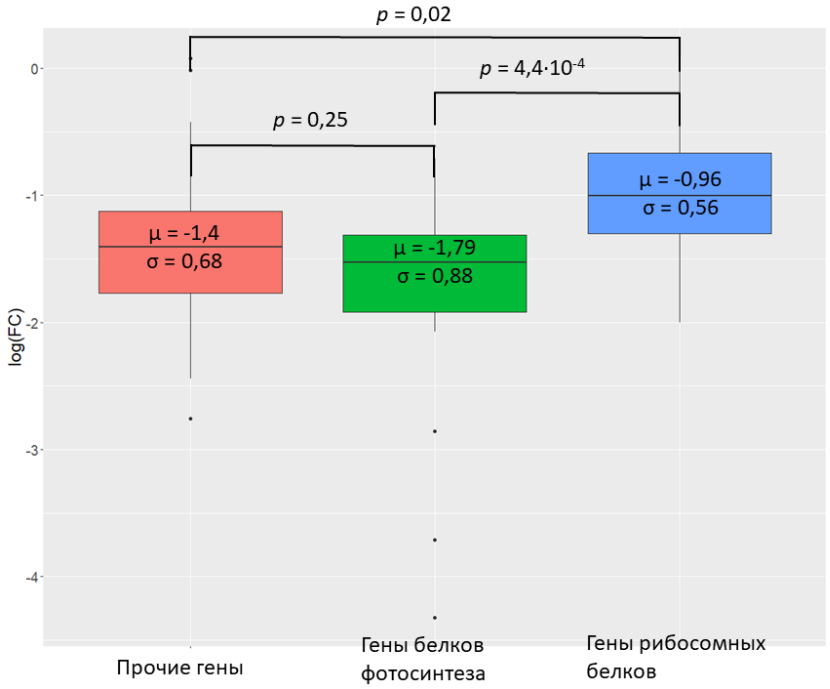

Рис 18. Изменение экспрессии разных функциональных групп пластидных генов в линии i:Bw*Blp* по сравнению с линией Bowman

Достоверность различий, оценённая с помощью теста Манна-Уитни, между группами «гены фотосинтеза» и «прочие гены» составляет 0,2524, между группами «гены рибосомных белков» и «прочие гены» – 0,02; наконец, между группами «гены фотосинтеза» и «гены рибосомных белков» – $4{,}38\ 10^{-4}$. Различия между экспрессией генов, связанных с фотосинтезом и генов, не связанных напрямую с фотосинтезом или функцией рибосом,

нельзя считать достоверными. Однако, различия в экспрессии генов фотосинтеза и генов, связанных с функцией рибосом, достоверны.

3.2.7 Экспрессия генов района Blp

Отдельно была рассмотрена экспрессия 21 гена, входящего в район Blp. Для 7 из числа этих генов наблюдается экспрессия выше порогового значения. При этом, достоверное изменение экспрессии наблюдается только для одного из генов, входящих в этот район. Это ген HORVU1Hr1G087010, кодирующий фосфатазу пурпурной кислоты. Он повышает свою экспрессию в линии i:Bw*Blp* в 39,75 раз ($\log_2(FC) = 5{,}313$) с достоверностью $p=1{,}55\cdot10^{-4}$ после поправки на множественное сравнение.

3.2.8 Реконструкция транскриптома *de novo*

Для библиотек RNA-seq из эксперимента на линиях i:Bw*Blp* и Bowman были построены индивидуальные сборки транскриптома *de novo* и мета-сборка транскриптома, после чего было проведено сравнение качества полученных сборок. В таблице 16 приведены основные параметры сборок транскриптома.

Таблица 16. Характеристики индивидуальных сборок *de novo* и мета-сборки траснкриптома леммы ячменя линий i:Bw*Blp* и Bowman

Сборка	Размер, контигов		N50	Средняя длина	Прочтений картировано, %
	Избыточная	Неизбыточная			
Abyss	214465	34987	606	490,32	68,75
Spades	31453	24401	1046	824,6	58,25
Trinity	116897	34363	891	661,59	66,55
GG	122304	39319	976	707,83	77,55
Мета-сборка	133070	32466	1056	775,73	72,07

Исходная избыточная мета-сборка транскриптома i:Bw*Blp* и Bowman состоит из 133070 контигов. После удаления избыточности мета-сборка состоит из 32466 контигов суммарной длиной 25184753 оснований. Таким образом, в ходе удаления избыточности

количество контигов было уменьшено до 24,4% от исходного. Также, отметим, что мета-сборка транскриптома в данном эксперименте имеет более высокое значение длин контигов N50, чем индивидуальные сборки, из которых она составлена. 72,1% всех прочтений из библиотек эксперимента было картировано на мета-сборку транскриптома. По этому показателю мета-сборка уступает сборке GG (77,6%), но опережает три другие индивидуальные сборки.

Был проведён поиск известных CDS в сборке транскриптома *de novo* исследуемых линий с помощью программы TransRate. Результаты поиска приведены в таблице 17. Как можно видеть из таблицы 17, от 19848 контигов в сборке Spades до 29412 контигов в сборке GG обнаруживают гомологию к известным CDS ячменя. При этом, наибольшее количество CDS ячменя обнаружено в сборке Trinity. Однако, наибольшее количество CDS ячменя, покрытых контигами сборки не менее чем на 95% своей длины, обнаружены в мета-сборке транскриптома – 1825 CDS. Доля контигов из сборки, для которых была обнаружена гомология к известным CDS ячменя, в мета-сборке составляет 74,5%, что ниже, чем у всех индивидуальных сборок, кроме сборки Trinity.

Таблица 17. Количество CDS ячменя, обнаруженных в *de novo* сборках транскриптома в эксперименте с линией i:Bw*Blp* и Bowman

Сборка	контигов найдено	% контигов	CDS найдено	p_95
Abyss	25804	0,738	18981	1224
Spades	19848	0,813	16818	1017
Trinity	22793	0,663	21885	1478
GG	29412	0,748	19947	1597
Мета-сборка	24194	0,745	19665	1825

Далее был проведён поиск перекрывания полученных для индивидуальных сборок транскриптома списков CDS, и была произведена оценка вклада каждой из индивидуальных сборок в общую структур мета-сборки. Перекрывание таких множеств показано на рисунке 19.

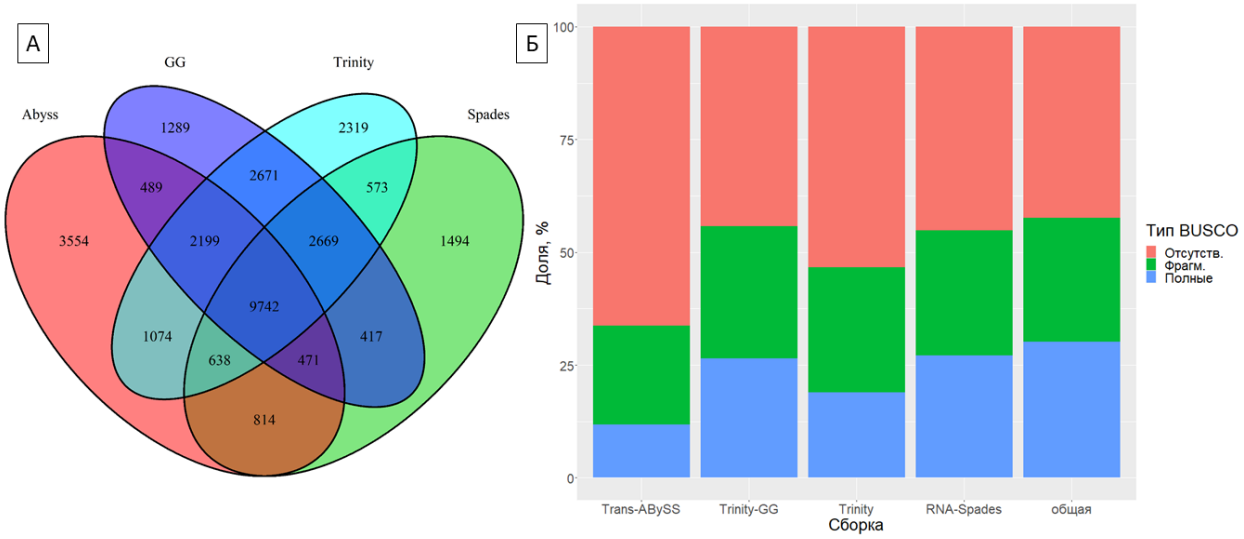

Рис. 19. (А) Пересечение множеств CDS, обнаруженных в индивидуальных сборках транскриптома *de novo* эксперимента Blp. (Б) Значения критерия BUSCO для разных сборок транскриптома в эксперименте Blp

Во всех четырёх индивидуальных сборках транскриптома *de novo* были обнаружены 9742 CDS. Уникальными для какой-либо из индивидуальных сборок *de novo*, то есть обнаруженными только в этой сборке и ни в одной из остальных, были 8656 CDS. Из этих уникальных для разных сборок CDS наибольшее количество – 3554 были уникальными для сборки abyss, наименьшее количество – 1289 были уникальными для сборки GG. Наибольшее количество общих CDS имеют сборки GG и Trinity – 17281 CDS были обнаружены в обеих этих сборках.

Суммарно 57,6% всех последовательностей BUSCO из набора для покрытосеменных организмов встречаются в избыточной мета-сборке транскриптома – больше, чем у либой из индивидуальных сборок. В полном виде представлены 30,2% последовательностей, в фрагментированном – 27,4%.

Из мета-сборки транскриптома 16813 транскриптов обладают уровнями экспрессии, достаточными для идентификации дифференциальной экспрессии. Достоверное понижение экспрессии было определено для 470 контигов, повышение – для 848 контигов.

В мета-сборке обнаружено 293 новых контигов. Из них 65 и 116 были удалены как артефакты сборки *de novo* и контаминация чужеродным материалом, соответственно. Оставшиеся 112 были проанализированы далее. Два контига достоверно повышают

экспрессию в линии i:Bw*Blp,* ещё два – понижают. Для этих контигов были рассмотрены гомологи из базы данных NCBI Protein nr и обнаруженная доменная структура. Эти контиги перечислены в таблице 18.

В контиге DN12020 не было обнаружено известных белковых доменов с помощью как NCBI Structure, так и InterProScan. Кодирующий потенциал контига оценён как 0,02. В контигах DN16564 и DN19100 были обнаружены с помощью NCBI Structure каталитические домены серин/треониновой фосфатазы ($E = 1{,}03 \cdot 10^{-10}$) и протеин-киназы ($E = 7{,}27 \cdot 10^{-44}$), соответственно. InterProScan определил в аминокислотном продукте контига DN16564 фосфатаза-подобный домен типа PPM (IPR001932) и отнёс аминокислотный продукт к семейству фосфатаза-подобных белков типа PPM (IPR036457). В аминокислотном продукте контига DN19100 были определены с помощью InterProScan каталитический домент тирозин-киназы (IPR020635), каталитический домент серин-треониновой/тирозиновой киназы (IPR001245) и домен протеинкиназы (IPR000719), а также активный сайт серин/треониновой киназы (IPR008271). Этот аминокислотный продукт был отнесёт к суперсемейству киназа-подобных белков (IPR011009). Кодирующий потенциал контигов DN16564 и DN19100 был оценён как 0,036 и 0,056, соответственно. Однако, сервис cpc2 определил размеры их аминокислотных продуктов как 80 и 110 аминокислотных остатков, соответственно, в то время как EvidentialGene определяет размеры их размеры как 121 и 140 аминокислотных остатков, соответственно.

Табл. 18. Новые контиги, обнаруженные в сборке транскриптома ячменя линии i:Bw*Blp*

Контиг	Диф. Экспр.		Лучший гомолог из базы NCBI Protein		
	logFC	FDR	идентификатор	описание	E-критерий
DN12020	-1,83	0,0036	CDY21223.1	BnaC03g54010D, *Brassica napus*	4,74
DN16564	1,24	0,026	XP_044946327.1	Предположительная фосфатаза белков 2C 68, *H. vulgare*	$4{,}14 \cdot 10^{-33}$
DN19100	2,82	0,0017	XP_044955417.1	Предположительная серин/треонин киназа белков PBL15, *H. vulgare*	$1{,}8 \cdot 10^{-75}$
DN21394	-1,03	0,0058	XP_037411011.1	Цитохром P450 709B1-подобный, *T. dicoccoides*	0

В пептидном продукте контига DN21394 был обнаружен домен цитохрома P450 (E = 0,0) как с помощью сервиса NCBI Structure, так и в результате работы сервиса InterProScan (IPR001128). Этот сервис также отнёс данный белковый продукт к семейству цитохромов P0450, E-классу, группе I (IPR002401). Для этого контига был определён кодирующий потенциал, равный 0,34. Однако, при этом размер его пептидного продукта был определён как 102 аминокислоты; программы из линейки EvidentialGene определяли размер его пептидного продукта как 523 аминокислоты. Идентификация доменной структуры и поиск гомологов также проводились именно для белкового продукта длиной в 523 аминокислоты.

Глава 4. Обсуждение

4.1 Методология обработки данных RNA-seq

4.1.1 Оценка качества библиотек коротких прочтений

В данной работе был проведён анализ библиотек коротких прочтений, полученных путём секвенирования транскриптома леммы ячменя двух изогенных линий и сорта Bowman на платформе IonTorrent PGM. Данные, получаемые с помощью секвенаторов IonTorrent, имеют свою специфику, которая отличает их от данных, полученных на платформах линейки Illumina, являющихся на данный наиболее распространёнными секвекнаторами второго поколения. В частности, платформа IonTorrent даёт достаточно длинные (150-200 нуклеотидов) прочтения, однако глубина секвенирования оказывается несколько ниже, чем у таких платформ, как Illumina NextSeq 550 или Illumina HiSeq.

В данной работе глубина секвенирования транскриптома леммы ячменя составляет 22,9 миллионов прочтений со средней длиной 175 нуклеотидов в эксперименте с линией i:Bw*Alm*, и 12,9 миллионов прочтений со средней длиной 183 нуклеотида в эксперименте с линией i:Bw*Blp*. И в том, и в другом эксперименте использовано по шесть библиотек коротких прочтений. Таким образом, средний размер очищенной библиотеки составляет 3,81 млн. прочтений в эксперименте с линией i:Bw*Alm* и 2,15 млн прочтений в эксперименте с линией i:Bw*Blp*. Такой глубины секвенирования достаточно для определения экспрессии генов и поиска дифференциальной экспрессии [Sims и др., 2014], а также *de novo* реконструкции транскриптома [Patterson и др., 2019].

В то же время, этой глубины секвенирования недостаточно для точного определения дифференциальной экспрессии изоформ [Katz и др., 2010] и поиска полиморфизмов [Quaglieri, Flensburg, Speed, 2019]. Более того, считается, что для точного определения дифференциальной экспрессии изоформ лучше подходят парные прочтения [Williams и др., 2014], в отличие от использованных в данном исследовании одиночных прочтений.

Учитывая, что данная работа проводится на материале почти изогенных линий ячменя, разработанных на основе сорта Bowman и отличающиеся от него только отдельными участками хромосом 3H и 1H, наличия полиморфизмов можно ожидать только в транскриптах генов, локализованных в этих участках. Кроме того, в отличие от данных геномного секвенирования, секвенирование транскриптомов даёт неравномерное

покрытие, пропорциональное уровням экспрессии генов [Jehl и др., 2021]. Следовательно, достоверное определение полиморфизмов возможно только в тех генах, которые имеют достаточно высокие уровни экспрессии. В случае двух изучаемых почти изогенных линий ячменя, это сужает область поиска полиморфизмов до тех генов, которые локализованы в отличающихся участках хромосом и имеют высокий уровень экспрессии, что в результате даёт крайне малое количество генов.

Поиск полиморфизмов в транскриптомах исследуемых линий не входит в задачи данной работы. Для решения же поставленных задач – определения дифференциальной экспрессии генов и *de novo* реконструкции транскриптома – глубина секвенирования является достаточной.

4.1.2 Сравнение конвейеров биоинформатической обработки данных

В данной работе был предложен метод сравнения конвейеров обработки данных RNA-seq, позволяющий подобрать оптимальный для имеющихся данных конвейер обработки. Этот метод учитывает четыре показателя, первые два из которых – доля картированных прочтений библиотек и доля уникально картированных прочтений библиотек – характеризуют качество процедуры картирования. Два других показателя – коэффициент корреляции Пирсона между уровнями изменения экспрессии, идентифицированными в результате анализа данных RNA-seq и определённых с помощью количественной от-ПЦР для ряда контрольных генов, и среднестатистическое отклонение такого коэффициента, полученное на подвыборках контрольных генов – характеризуют точность определения дифференциальной экспрессии генов.

В данной работе предлагается проводить приоретизацию конвейеров биоинформатической обработки данных RNA-seq основываясь на сумме рангов приведённых выше четырёх параметров. Авторы не утверждают, что данный метод приоретизации конвейеров является единственно возможным, и допускают, что в дальнейшем этот метод может быть усовершенствован. К примеру, использование взвешенных рангов может улучшить процедуру выбора конвейеров и повысить точность определения дифференциальной экспрессии генов. Однако, авторы подчёркивают необходимость использования множественных конвейеров и их последующей приоретизацией для получения наиболее точных результатов.

Многие другие авторы также указывают на необходимость выбора конкретных программ для биоинформатической обработки данных RNA-seq [Conesa и др., 2016; Rajkumar и др., 2015; Williams и др., 2016]. В данной работе была проанализирована производительность 36 конвейеров на двух наборах данных, и с помощью предложенного метода приоретизации выбраны конвейеры, наиболее оптимальные для каждого из этих двух наборов данных. Авторы выражают надежду, что эти результаты в дальнейшем будут использованы в других работах по анализу данных RNA-seq и помогут улучшить получаемые результаты, сделав их более точными и надёжными.

4.1.3 Сравнение конвейеров *de novo* реконструкции транскриптома

В данной работе была проведена *de novo* реконструкция транскриптома леммы ячменя двух почти изогенных линий, контрастных по окраске колоса. Были использованы три программы для реконструкции транскриптома *de novo*: Trans-ABySS, RNA-Spades и Trinity. Кроме того, программа Trinity была использована в режиме ‘genome-guided’, в котором программа реконструирует транскрипты, основываясь на картировании библиотек коротких прочтений на референсный геном организма. Таким образом, для каждой из двух линий были получены по четыре сборки транскриптома. Далее, для каждой из линий полученные сборки были объединены в один общий набор транскриптов, из которого затем была удалена избыточность с помощью скриптов из линейки EvidentialGene. Полученные в результате не-избыточные наборы транскриптов для каждой из линий будем называть мета-сборкой транскриптома данной линии.

Для каждой из двух линий полученные сборки транскриптома и составленные из них мета-сборки были сравнены по ряду показателей, для которых была проведена нормировка по методу, предложенному в работе [Hölzer, Marz, 2019] (подробнее см. раздел 2.2.2). В результате было показано, что для обеих линий мета-сборки транскриптома опережают любые из индивидуальных сборок по совокупности оцененных параметров. При этом, однако, мета-сборки могут уступать тем или иным индивидуальным сборкам по отдельным оцененным показателям.

Авторы не утверждают, что предложенный в данной работе набор оценённых параметров сборок транскриптома является единственно верным. Также, авторы не утверждают, что метод нормализации оценённых показателей [Hölzer, Marz, 2019],

адаптированный в данной работе, является единственно верным. Однако, авторы заостряют внимание на проблеме, поднимаемой во многих работах, посвящённых анализу методологии *de novo* реконструкции транскриптома из данных секвенирования второго поколения. Авторы надеются, что метод создания мета-сборки транскриптома и последующего сравнения качества мета-сборки и индивидуальных сборок транскриптома, описанные в данной работе, позволят в дальнейшем повысить качество реконструируемых *de novo* транскриптомов в последующих исследованиях.

4.2 Транскриптомный анализ линии i:Bw*Alm*

4.2.1 Функциональный анализ дифференциально экспрессирующихся генов

Для библиотек RNA-seq транскриптома линии i:Bw*Alm* и сорта Bowman, взятого в качестве контроля, был выявлен наиболее оптимальный конвейер для биоинформатической обработки имеющихся данных секвенирования. Этот конвейер состоит из следующих стадий: фильтрация исходных библиотек коротких прочтений с помощью программы Prinseq и удаления рибосомных РНК путём картирования библиотек на последовательности некодирующих РНК ячменя; картирования прочтений на референсный геном с помощью программы Dart; поиска дифференциальной экспрессии генов с помощью пакета EdgeR для языка R. С помощью такого конвейера была достигнута наибольшая корреляция значений изменения экспрессии, определённых с помощью анализа данных RNA-seq с наблюдаемыми экспериментально с помощью количественной полимеразной цепной реакции.

Сравнительный анализ транскриптомов почти изогенной линии i:Bw*Alm* и сорта Bowman показал значительное расхождение в регуляции экпрессии генов – в то время как 1287 генов понижают экспрессию в линии i:Bw*Alm,* только 78 генов повышают экспрессию в этой линии. Такое резкое смещение уровней изменения экспрессии в сторону понижения в клетках леммы растений этой линии по сравнению с контролем связано со спецификой наблюдаемого физиологического процесса. В клетках этих тканей отсутствует хлорофилл, в связи с чем не происходит фотосинтез. Фотосинтез – важнейший для жизнедеятельности растений процесса, в которые вовлечено огромное количество генов [Dubreuil и др., 2018]. Большое количество генов кодируют белки, участвующие в биосинтезе хлорофилла, сборке фотосистем. Следовательно, можно предположить, что в клетках, в которых эти процессы

не происходят, экспрессия этих генов должна быть понижена. Это объясняет большее количество генов со сниженной экспрессией у i:Bw*Alm* генов, нежели генов с повышенной экспрессией у этой линии.

Функциональный анализ ДЭГ выявил участие генов, имеющих пониженную в линии i:Bw*Alm* экспрессию, в одиннадцати метаболических путях. Для этих генов значимо обогащены 234 термина генной онтологии. Рассмотрим эти метаболические пути и термины генной онтологии подробнее.

Среди метаболических путей можно выделить пути, непосредственно связанные с фотосинтезом. В эту группу входят цикл Кальвина-Бенсона, световая фаза фотосинтеза и шунт РБФК. Функциональная связь этих метаболических путей и формирования частично бесхлорофилльного фенотипа представляется достаточно очевидной. В самом деле, метаболический путь «Световая фаза фотосинтеза» связан с фотосинтезом по определению. Цикл Кальвина-Бенсона также тесно сопряжён с фотосинтезом, поскольку именно энергия, полученная в ходе световой фазы фотосинтеза, уходит на фиксацию атмосферного углекислого газа [Lu, Yao, 2018]. В ходе одной из реакций цикла Кальвина-Бенсона происходит присоединение молекулы углекислого газа в виде карбоксильной группы к рибулозо-1,5-бисфосфату с образованием 3-фосфо-D-глицерата [Berry, Mure, Yerramsetty, 2016]. Эта реакция катализируется ферментом рибулозо-1,5-бисфосфаткарбоксилазой/оксигеназой (РБФК, RuBisCO). Этот фермент также участвует в метаболическом пути «Шунт РБФК». В ходе этого пути b-D-фруктофураноза-6-фосфат в результате каскада ферментативных реакций преобразуется в ацетил-CoA [Schwender и др., 2004].

Обогащение многих терминов генной онтологии также очевидно связано с особенностями наблюдаемого фенотипа. Сюда можно отнести термины из категории «Биологический процесс»: «фотосинтез», «фотосинтетический транспорт электронов в фотосистеме I» и «фотосинтетический транспорт электронов в фотосистеме II», «стабилизация фотосистемы II», «биосинтез хлорофилла». Термины из категории «Клеточная локализация», относящиеся к этой группе: «хлоропласт», «реакционный центр фотосистемы II», «водоокисляющий комплекс фотосистемы II» и «фотосистема I». Наконец, следующие термины генной онтологии из категории «Молекулярная функция» относятся к этой группе: «рибулозобисфосфаткарбоксилазная активность», «путь

циклического переноса электронов при фотосинтетической активности», «связывание хлорофилла».

Другая группа метаболических путей, которую можно выделить – пути, связанные с усвоением азота. В эту группу входят метаболические пути «усвоение аммония», «биосинтез L-глутамина I», «биосинтез L-глутамина III», и «ассимиляционное усвоение нитратов». На первый взгляд, связь наблюдаемого частично-альбиносного фенотипа ячменя и таких процессов, как усвоения аммония или нитратов, кажется неясной. Процессы же биосинтеза глутамина и аденозиновых нуклеотидов, как мы увидим далее, связаны с усвоением аммония.

Растения в основном усваивают азот из почвы, впитывая его корнями вместе с водой в виде растворимых солей нитратов или аммония [Baslam и др., 2020]. Нитраты сперва восстанавливаются до нитритов ферментом нитрат-редуктазой, а затем с помощью фермента нитрит-редуктазы – до аммония [Schläpfer и др., 2017]. Аммоний же участвует в реакциях, катализируемых ферментами глутаминсинтетазой (GS; EC 6.3.1.2) и глутаматсинтазой (Глутамин-2-оксоглутаратаминотрансфераза, GOGAT). Первый из этих ферментов катализирует присоединение аммония к глутамату с образованием глутамина [Miflin, Habash, 2002]; второй катализирует перенос амино-группы с молекулы глутамата на молекулу 2-оксоглутарата, в результате чего образуются две молекулы глутамина [Bao и др., 2015, с. 201]. В результате всех этих процессов неорганический азот, усвоенный растениями из почвы, входит в состав органических соединений.

Но всасывание минеральных солей – не единственный процесс, в результате которого ионы аммония появляются в клетках растений. Помимо этого, аммоний образуется при метаболизме аминокислот, азотистых оснований и других соединений, а также в ходе каскада ферментативных реакций, известного как фотодыхание [Betti и др., 2016]. Именно в результате фотодыхания в клетках растений появляется наибольшее количество аммония; при фотодыхании может выделяться до десяти раз больше аммония, чем растения могут получить в ходе всасывания аммония и нитратов корнями [Keys и др., 1978]. Таким образом, хотя фотодыхание и не приводит к фиксации неорганического азота из почвы или атмосферы, этот процесс играет важную роль в круговороте азотсодержащих молекул в клетках растения. Метаболический путь фотодыхания и сопутствующие процессы схематично показаны на рисунке 20.

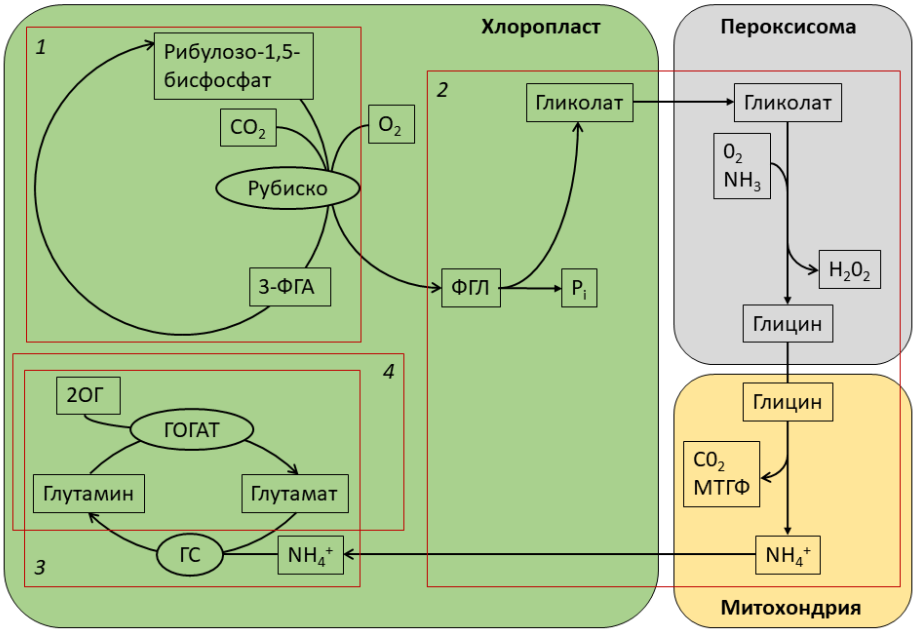

Рис. 20. Метаболические пути цикла Кальвина, фотодыхания и усвоения аммония [Bauwe, Hagemann, Fernie, 2010]. 3-ФГА – 3-фосфоглицеральдегид, ФГЛ – фосфоглицерат, МТГФ – метилентетрагидрофолат, 2-ОГ – 2-оксоглутарат, ГС – глутаминсинтаза, ГОГАТ – глутаматсинтетаза. Представлены метаболические пути, аннотированные в базе данных PlantCyc: (1) – цикл Кальвина-Бенсона; (2) – фотодыхание; (3) – цикл усвоения аммония; (4) – усвоение аммония II.

Фотодыхание неразрывно связано с циклом Кальвина-Бенсона [Timm и др., 2016]. В цикле Кальвина-Бенсона, как описано выше, происходит усвоение неорганического углерода за счёт действия фермента РБФК. Однако, этот фермент способен катализировать не только карбоксилирование рибулозо-1,5-бисфосфата, но и его оксигенирование, в результате чего образуются 3-фосфо-D-глицерат и 2-фосфогликолат [Hagemann, Bauwe, 2016]. Если первое из этих двух соединений далее снова используется в цикле Кальвина-Бенсона, то второе метаболизируется именно в ходе фотодыхания [Peterhansel, Maurino, 2011]. В качестве побочных продуктов при фотодыхании выделяются пероксид водорода, который затем метаболизируется каталазами, углекислый газ и аммоний [Peterhansel и др., 2010].

Цикл Кальвина-Бенсона, как уже было сказано выше, тесно сопряжён с фотосинтезом. Следовательно, в отсутствие хлорофилла и при невозможности фотосинтеза, реакции цикла Кальвина-Бенсона также останавливаются, а вместе с ними и фотодыхание [Timm и др., 2016]. В этом случае огромное количество аммония, которые выделяется при фотодыхании, не поступает в клетки. Этим объясняется понижение экспрессии генов, входящих в метаболические пути усвоения аммония, которое наблюдается в данной работе.

Что касается метаболического пути «синтез аденозиновых нуклеотидов», считается, что этот метаболический путь у растений локализован в пластидах [Witte, Herde, 2020]. Аденин используется в огромном количестве различных метаболических реакций [Haferkamp, Fernie, Neuhaus, 2011], но кроме того, путь биосинтеза этого нуклеотида служит для утилизации аммония в клетках растений [Smith, Atkins, 2002]. Как уже было сказано ранее, в клетках ячменя линии i:Bw*Alm*, утилизация аммония в целом подавлена по сравнению с ячменём сорта Bowman. По-видимому, именно этим объясняется наблюдаемое понижение экспрессии генов, входящих в этот метаболический путь, в лемме растений данной линии.

Третья группа метаболических путей связанна с процессом аэробного дыхания: «аэробное дыхание I», называемое также «цитохромный путь аэробного дыхания» и «аэробное дыхание III», называемое также «альтернативный оксидативный путь аэробного дыхания». Как мы далее увидим, с этими метаболическими путями тесно связан ещё один значимо обогащённый ДЭГ метаболический путь, «фосфорилирование и дефосфорилирование NAD/NADH».

Аэробное дыхание у эукариот происходит в митохондриях [Osellame, Blacker, Duchen, 2012]. Митохондрии, как и пластиды, являются полуавтономными органоидами, имеющими свой геном и окружёнными двойной мембраной [McCarron и др., 2013]. В ходе аэробного дыхания высокоэнергетические электроны из молекул убихинола, сукцината и NADH переносятся по так называемой дыхательной цепи, одновременно с чем происходит перенос протонов из матрикса в межмембранное пространство митохондрий ферментативными комплексами I, III и IV [Schertl, Braun, 2014]. Ферментативный комплекс II не переносит протоны через внутреннюю мембрану, но синтезирует убихинол, который затем используется комплексом III [Huang, Millar, 2013]. В результате создаётся протонный градиент между матриксом и межмембранным пространством митохондрий. Затем

протоны движутся по градиенту концентрации через мембранный канал фермента АТФ-синтазы, которая использует полученную от протонов энергию для фосфорилирования АДФ [Braun, 2020]. Таким образом, энергия, запасённая в электронах, содержащихся в NADH и в сукцинате, в конечном итоге переводится в энергию АТФ. Этот процесс также называется цитохромным дыханием, так как одним из промежуточных соединений в цепи переноса электронов является цитохром *c* [Siedow, Umbach, 2000].

Функции митохондрий и пластид у растений тесно координированы друг с другом [Toshoji и др., 2012]. В частности, известна координация именно таких процессов, как фотосинтез и аэробное дыхание [Yoshida, Terashima, Noguchi, 2007]. Однако, у растений, помимо цитохромного аэробного дыхания, существует так называемый альтернативный оксидативный путь аэробного дыхания [Siedow, Umbach, 2000]. Этот путь также использует энергию, запасённую в NADH, и переведённую затем в форму убихинола, но не создаёт межмембранный протонный градиент. Напротив, этот процесс переносит электроны напрямую в молекулярный кислород, что прерывает дыхательную цепь [Schertl, Braun, 2014]. Это происходит за счёт действия фермента, известного как альтернативная оксидаза, AOX [Siedow, Umbach, 2000]. Этот фермент обнаружен у всех известных растений, а также у некоторых грибов и у очень небольшого количества животных [Matus-Ortega и др., 2011].

В базе данных PlantCyc, которая была использована для определения участия генов ячменя в метаболических путях, пути цитохромного аэробного дыхания и альтернативного оксидативного аэробного дыхания содержат 242 и 184 гена, соответственно. При этом, в пути альтернативного оксидативного дыхания приведены не только гены, кодирующие непосредственно альтернативную оксидазу, но и гены, кодирующие ферментативные комплексы I и II из цепи переноса электронов, которые синтезируют убихинол, используемый в качестве субстрата альтернативной оксидазой. Ферментативные комплексы в мембране митохондрий и перенос электронов показаны на рисунке 21.

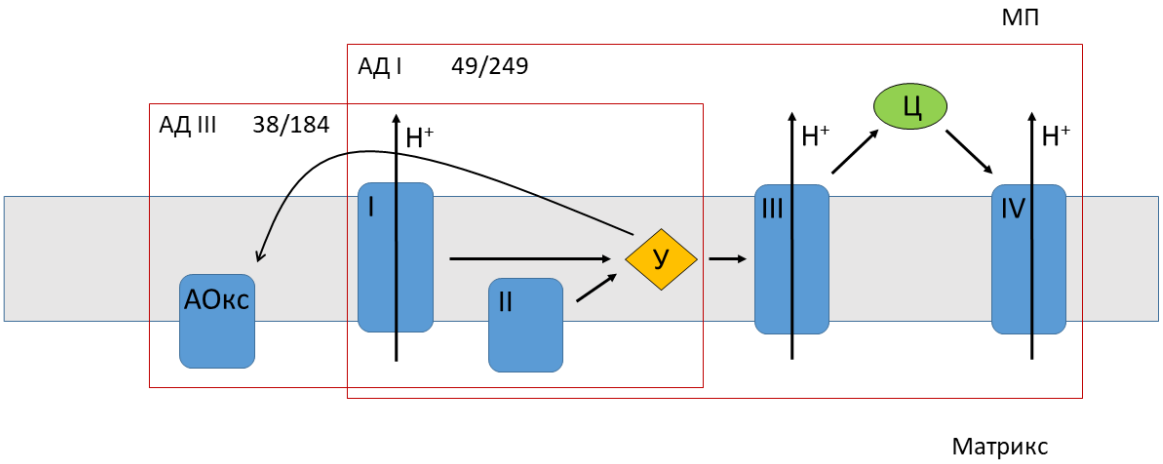

Рис. 21. Цепь переноса электронов в митохондриях [https://pmn.plantcyc.org/BARLEY/]. У – убихинол, Ц – цитохром, АОкс – альтернативная оксидаза, I-IV – ферментативные комплексы I-IV соответственно. Выделены аннотированные в базе данных PlantCyc метаболические пути: АДI – Аэробное дыхание I, АДIII – Аэробное дыхание III. Для этих метаболических путей приведено количество участвующих в них ДЭГ относительно общего количества участвующих генов.

В данном исследовании наблюдается понижение экспрессии 38 генов, входящих в путь альтернативного оксидативного дыхания. Все эти гены также входят в путь цитохромного дыхания. Помимо этих 38 ДЭГ, ещё 11 генов, входящих в путь цитохромного дыхания, понижают свою экспрессию, в том числе 1 ген, локализованный в геноме пластид.

Метаболический путь «фосфорилирование и дефосфорилирование NAD/NADH», приведённый в базе данных PlantCyc, откуда были взяты данные об участии генов ячменя в метаболических процессах, включает в себя несколько реакций для фосфорилирования и дефосфорилирования отдельно NAD и отдельно NADH. Эти две части метаболического пути связаны процессом взаимного превращения NAD и NADH. Белки, участвующие в этих взаимных превращениях, совпадают с белками, составляющими комплекс I из цепи переноса электронов в аэробном дыхании.

В данной работе наблюдается понижение экспрессии 38 генов, входящих в путь «фосфорилирование и дефосфорилирование NAD/NADH». Этот список ДЭГ полностью совпадает со списком ДЭГ, входящих в путь «аэробное дыхание III». Все эти ДЭГ также входят в путь «аэробное дыхание I». Таким образом, функциональная связь

фосфорилирования и дефософорилирование NAD/NADH с исследуемым фенотипическим проявлением объясняется частичным перекрыванием генов, входящих в этот метаболический путь, с генами аэробного дыхания у растений.

Дифференциальной экспрессии генов, кодирующих альтернативную оксидазу, фермент, специфичный для альтернативного оксидативного дыхания, не наблюдается. Это согласуется с имеющимися данными, такими, как сообщение, что в белых секторах листьев мутантов *Arabidopsis* не повышается содержания белка AOX, в то время как в зелёных секторах листьев этого мутанта, страдающих от фотооксидативного стресса, существенно повышается содержание этого белка [Yoshida, Terashima, Noguchi, 2007].

Гены с повышенной экспрессией в линии ячменя i:Bw*Alm* участвуют в 15 метаболических путях; однако, ни для одного из этих путей обогащение ДЭГ не оказалось статистически значимым. Четыре термина генной онтологии оказались значимо обогащены для генов с повышенной экспрессией у линии i:Bw*Alm*. Три из этих терминов связаны с реакцией на стресс, четвёртый термин – протеолиз.

Детально были рассмотрены гены, локализованные в районе Alm. Ранее было показано, что десять линий ячменя, демонстрирующих фенотип, присущий исследуемой линии i:Bw*Alm*, имеют различные мутации в гене HORVU3Hr1G032440, локализованном в районе Alm [Taketa и др., 2021]. Этот ген кодирует транскрипционный фактор Golden-like2, связанный в развитии хлоропластов в плодах растений [Nguyen и др., 2014]. Это делает данный ген перспективным кандидатом на роль гена *Alm*. В настоящей работе, однако, не наблюдается изменения экспрессии этого гена между линией i:Bw*Alm* и сортом Bowman. Таким образом, данных для подкрепления гипотезы Taketa и соавторов [2021], что ген HORVU3Hr1G032440 является геном *Alm*, недостаточно.

Из семи определённых в настоящей работе ДЭГ, локализованных в районе Alm, один повышает экспрессию в линии i:Bw*Alm* по сравнению с сортом Bowman. Это ген HORVU3Hr1G034230, кодирующий 40S рибосомный белок. У растений, как и у других эукариот, представлены как цитоплазматические 90S рибосомы, так и митохондриальные 70S рибосомы; но в отличие от других эукариот, у растений представлены также пластидные рибосомы [Romani и др., 2012]. Описаны мутанты по генам белков пластидных рибосом, отличающиеся полным или частичным альбизинмом [Sáez-Vásquez, Delseny, 2019; Zhou и др., 2021], в том числе, летальные [Lee и др., 2019]. В данном случае, однако,

наблюдается изменение экспрессии гена, кодирующего белок, входящий в состав эукариотической 90S рибосомы. Гены, кодирующие белки цитоплазматических рибосом, представлены семействами паралогов [Martinez-Seidel и др., 2020], которые экспрессируются и участвуют в составлении рибосом в органо- и тканеспецифичном паттерне [Hummel и др., 2015]. Также, разные паралоги тех или иных рибосомных генов по-разному экспрессируются в условиях стрессов [Eskelin и др., 2019]. По-видимому, именно этим объясняется повышение экспрессии данного гена в линии i:Bw*Alm*. Однако, насколько авторам известно, до сих пор не было описано повышения экспрессии генов, кодирующих белки цитоплазматических рибосом, при частичном альбинизме растения.

Рассмотрим подробнее гены, локализованные в районе Alm и понижающие экспрессию в линии i:Bw*Alm* (См. табл. 10). Ген HORVU3Hr1G034100, кодирующий белок из суперсемейства купин-доменных белков, понижает экспрессию примерно в 1,8 раз. Купины – семейство белков, характеризующихся наличием консервативного β-цилиндра в третичной структуре [Dunwell, Purvis, Khuri, 2004]. Это семейство отличается огромным разнообразием функций [Dunwell, Purvis, Khuri, 2004]. В числе этих функций – запасающая, присущая купинам, встречающимся в семенах растений [Dunwell и др., 2001], некоторые из которых способны связывать молекулы олигосахаридов [Uberto, Moomaw, 2013]. В целом, для этого гена наблюдается достаточно слабое понижение уровней экспрессии, и значение статистической достоверности изменения экспрессии близко к пороговому (0,024), поэтому связь данного гена с изучаемым процессом не очевидна.

Гены HORVU3Hr1G034070 и HORVU3Hr1G034060, кодирующие гистоны H2A и H2B, понижают экспрессию в 50 и 68 раз, соответственно. Значения критерия достоверности *p* равны 0,0016 и 0,0013, соответственно. Гистоны считаются одними из самых стабильных белков, которые редко утилизируются клеткой; синтез новых гистонов происходит в основном в S-фазе клеточного цикла [Lyons и др., 2016]. Однако, гены, кодирующие каждый из типов гистонов (H2A, H2B, H3 и H4), представлены семействами паралогов [Jiang и др., 2020]. Несмотря на то, что эти паралоги высокогомологичны, методы RNA-seq могут различать их между собой и определять уровни экспрессии для отдельных конкретных генов, кодирующих гистоны [Iwaya и др., 2013]. Паралоги различаются паттернами экспрессии в разных тканях, органах и/или на разных стадиях развития [Kawashima и др., 2015]. Экспрессия атипичного гистона H3.15 у *A. thaliana* стимулирует

развитие каллуса [Yan и др., 2020]. Всё это позволяют предположить, что изменение экспрессии гистонов объясняется стрессовыми условиями, в которых находятся клетки леммы растений линии i:Bw*Alm* из-за неспособности фотосинтезировать. Однако, насколько авторам известно, до сих пор не было показано связи изменения экспрессии генов гистонов и формирования альбиносного фенотипа у растений.

Ген HORVU3Hr1G031940 кодирует протеазу, локализованную на внутренней мембране митохондрий. Экспрессия этого гена понижается в линии i:Bw*Alm* в 396 раз с достоверностью $1{,}27 \cdot 10^{-24}$. Протеазы, локализованные на внутренних мембранах митохондрий, отщепляют сигнальные пептиды от белков, транспортируемых в митохондрии из цитоплазмы [Horvath и др., 2015]. По некоторым оценкам, до трёх тысяч разных белков импортируются в митохондрии из цитоплазмы [Meyer, Letts, Maldonado, 2022]. Нарушение функций митохондриальных пептидаз, отщепляющих сигнальные пептиды, может приводить клетку к стрессовому состоянию [Horvath и др., 2015; Horvath и др., 2015]. Как уже было сказано, функции митохондрий и функции пластид в растительной клетке тесно связаны, следовательно, снижение экспрессии этого гена можно объяснить общим понижением транспорта белков из цитоплазмы в митохондрии.

Ген HORVU3Hr1G034640 кодирует транскрипционный фактор из семейства Myb/Myc. Этот ген понижает экспрессию в линии i:Bw*Alm* в 3,8 раз с достоверностью $2{,}02 \cdot 10^{-3}$. Физиологические процессы, регулируемые ТФ этого семейства, можно подразделить на четыре группы: (1) первичный и вторичный метаболизм; (2) клеточная пролиферация и идентификация; (3) развитие; (4) ответы на биотические и абиотические факторы окружающей среды [Dubos и др., 2010]. Как видно, это очень широкий спектр самых разнообразных процессов. Описано не менее 150 ТФ этого семейства у риса *Oryza sativa* L. и около 200 у арабидопсис [Katiyar и др., 2012]. В данном случае можно объяснить изменение экспрессии этого гена ответом на стресс, который растения линии i:Bw*Alm* испытывают в связи с отсутствием хлорофилла в органах, в норме способных к фотосинтезу. Однако, известно, что ТФ Golden2-like содержит MYB-ДНК связывающий домен [Zhao и др., 2021]. Этот ТФ контролирует развитие хлоропластов и синтез хлорофилла в зелёных органах растений [Safi и др., 2017], причём у двудольных растений этот ген имеет тканевую специфичность, и экспрессируется преимущественно в плодах [Nguyen и др., 2014]. И хотя тканевая специфичность работы этого гена, насколько авторам

известно, не была ранее показана для однодольных растений, это делает роль данного гена в развитии частично альбиносного фенотипа у растений линии i:Bw*Alm* интересной и заслуживающей дальнейшего более подробного изучения.

Наконец, ген HORVU3Hr1G032490 кодирует глюкоза-6-фосфат/фостфат транслокатор-подобный белок. Этот ген понижает свою экспрессию в 3,75 раз в клетках леммы растений линии i:Bw*Alm* с достоверностью $p = 7{,}25 \cdot 10^{-10}$. Глюкоза-6-фосфат/фосфат транслокатор переносит молекулу глюкоза-6-фосфат из цитоплазмы в пластиды с одновременным транспортом иона фосфата из пластид в цитоплазму [Flügge и др., 2003]. Этот белок в клетках растений локализован на мембранах пластид [Flügge и др., 2011], но появляются сообщения о возможной вторичной локализации этого белка на мембране пероксисом [Baune и др., 2020]. Нарушение экспрессии этого гена у арабидопсис замедляет прорастание семян и синтез хлорофилла в семядолях на свету [Dyson, Webster, Johnson, 2014]. Во взрослых растениях экспрессия наблюдается в основном в не-зелёных частях растений [Facchinelli, Weber, 2011], в особенности в клетках запасающих тканей [Wu, Wang, Zhang, 2021], но резко повышается экспрессия в листьях при повышении интенсивности освещения [Fabiańska, Bucher, Häusler, 2019]. Предполагается, что соотношение сахаров и фосфатов в пластидах и цитозоли при этом контролирует интенсивность фотосинтеза [Weise и др., 2019].

Из всего перечисленного видно, что роль этого гена в развитии растений и, вероятно, механизмах ответа на факторы внешней среды, до конца не ясна. Этот ген кажется возможным кандидатом на роль гена *Alm*, но делать выводы о его роли в формировании частично альбиносного фенотипа на данный момент затруднительно.

4.2.2 Анализ *de novo* реконструированного транскриптома

Была проведена реконструкция транскриптома линий i:Bw*Alm* и Bowman *de novo*, для чего были использованы несколько программ-сборщиков, с помощью которых были созданы индивидуальные сборки *de novo*, из которых впоследствии была скомпонована мета-сборка транскриптома. Было показано, что по совокупности ряда параметров, характеризующих полноту и качество сборки, мета-сборка транскриптома опережает индивидуальные сборки, из которых она была составлена. Это наблюдение согласуется с высказанными ранее наблюдениями о повышении качества реконструированного

транскриптома при использовании нескольких программ-сборщиков и дальнейшей компоновки результатов их работы в одну общую сборку [Cerveau, Jackson, 2016; Evangelistella и др., 2017].

При анализе реконструированного *de novo* транскриптома были обнаружены пять новых контигов, имеющих значимую дифференциальную экспрессию. Два из них – DN20386 и DN16917 – имеют лучшую гомологию к растительным последовательностям XP_044969885.1 и KAF6992501.1, соответственно, описанным как неохарактеризованный белок *H. vulgare* и гипотетический белок *T. aestivum*. Среди функционально охарактеризованных последовательностей эти контиги имеют лучшую гомологию к TTLL7-подобной тубулин полиглутамилазе *Acanthaster planci* и N-ацетилтрансферазе семейства GNAT *Planobispora takensis*, соответственно. Однако, уровень гомологии с этими последовательностям достаточно низок (E = 7,98 и 4,38). *P. takensis* – грам-положительная бактерия, *A. planci* – терновый венец, иглокожее семейства *Acanthasteridae* Sladen. Для этих последовательностей не было обнаружено известных белковых доменов. Кодирующий потенциал этих контигов оценён как достаточно низкий. В свете всего перечисленного видно, что на данный момент функционально охарактеризовать два данных транскрипта и высказать предположения об их роли в формировании изучаемого фенотипа ячменя не представляется возможным. Более того, нельзя утверждать наверняка, являются ли данные контиги транскриптами генов ячменя, контаминантами или же артефактами *de novo* сборки транскриптома. Для ответа на этот вопрос необходима независимая экспериментальная проверка с помощью полимеразной цепной реакции или иных молекулярно-генетических методов.

Аминокислотный продукт другого контига, GG_19862, имеет гомологию к серин/треониновой киназе белков RHS3 *A. tauschii*, злака, относящегося, как и ячмень, к трибе *Triticeae*. В нём также обнаружен каталитический домен протеинкиназы. Это указывает, что, по-видимому, данный контиг является транскриптом гена ячменя, кодирующего протеинкиназу. Протеинкиназы – ферменты, катализирующие присоединение фосфатной группы к молекулам белка. У растений физиологические функции этих ферментов обычно связывают с ответом на различные виды стресса, такие как холодовой [Guo, Liu, Chong, 2018], засуховый [Chen и др., 2021], биотический [Hake, Romeis, 2019]. Однако, появляются также сообщения об участии протеинкиназ в регуляции

клеточного цикла [Banerjee, Singh, Sinha, 2020] и регуляции размера семян у растений [Li, Xu, Li, 2019]. Ввиду такого многообразия функций этого класса ферментов, сложно делать выводы о конкретной роли данного гена в формировании частично альбиносного фенотипа ячменя.

Ещё один контиг, DN4161, содержит открытую рамку считывания, кодирующую аминокислотный продукт, гомологичный KH домен-содержащему белку PEPPER риса *O. sativa*. Обнаруженный в нём РНК-связывающий домен типа I K-гомологов подтверждает принадлежность аминокислотного продукта к этой группе белков. Домен такого типа впервые был описан в гетерогенном ядерном рибонуклеопротеине К человека [Siomi и др., 1993], следующие открытые подобные белки стали называть К-гомологами, K-homologs, KH. У животных такие белки участвуют в стабилизации транспортных РНК, регуляции транскрипции и транскрипционном сайленсинге [Yan и др., 2017]. У растений их функции изучены хуже [Yan и др., 2017]. В геноме арабидопсис обнаружено 30 генов, кодирующих KH-белки [Zhang и др., 2022], среди них выделяют связанные с ответом на тепловой стресс [Guan и др., 2013] и с регуляцией развития цветка [Rodríguez-Cazorla и др., 2015]. В частности, белок PEPPER, содержащий три KH-домена [Ripoll и др., 2006], опосредованно регулирует время цветения арабидопсис, участвуя в транскрипции и процессинге пре-мРНК гена *Flc*, одного из основных регуляторов времени цветения [Ripoll и др., 2009]. Помимо этого, у некоторых мутантов по гену *pep* наблюдается бледно-зелёный мезофилл листа [Ripoll и др., 2006]. В данной работе была идентефицирована значимая экспрессия транскрипта, кодирующего аминокислотный продукт, гомологичный этому гену, в линии i:Bw*Alm* и отсутствие его экспрессии в сорте Bowman. Всё это позволяет предположить участие этого гена в формировании исследуемого фенотипического проявления.

Наконец, транскрипт DN2647 содержит открытую рамку считывания, кодирующую аминокислотный продукт, имеющий в своём составе домен прохибитина и гомологию к прохибитину *S. pennellii*. Кодирующий потенциал этого транскрипта оценён как достаточно невысокий; однако, такая оценка основана на длине ОРФ, заниженной по сравнению с результатами, полученными с помощью EvidentialGene. Поиск гомологов и доменной структуры же проводился именно для пептидного продукта, полученного с помощью EvidentialGene. Таким образом, можно предположить, что кодирующий потенциал этого транскрипта достаточно высок.

Для этого белка обнаружена гомология с dao-5-подобным белком *T. dicoccoides*. Dao-5 – dauer and aged animal overexpression, белок *Caenorhabditis elegans*, участвующий в генезе ядрышка [Korčeková и др., 2012]. Мутанты *C. elegans* по этому гену имеют пониженную фертильность и нарушения в развитии половых органов [Lee и др., 2014]. Однако, учитывая наличие домена прохибитина в аминокислотном продукте транскрипта DN2647, представляется более вероятным принадлежность его именно к этому семейству белков.

Белки семейства прохибитинов достаточно консервативны и присутствуют в клетках всех эукариотических организмов [Artal-Sanz, Tavernarakis, 2009]. Внутри клеток прохибитины локализованы в митохондриях [Krupinska и др., 2020], где образуют комплексы из 12-16 белковых молекул. Однако, есть сообщения также и о ядерной локализации прохибитинов [Morrow, Parton, 2005; Wang и др., 2002].

Прохибитины участвуют в регуляции клеточного цикла [Nuell и др., 1991], переносе сигналов [Rajalingam, Rudel, 2005], старении и клеточной гибели [Morrow, Parton, 2005]. У растений прохибитины, помимо прочих функций, участвуют в регуляции развития различных органов и тканей [Huang, Yang, Zhang, 2019]. Прохибитины могут влиять на стабильность пластид [Chen и др., 2019], что косвенно указывает на их возможную связь с синтезом и накоплением хлорофилла. Прохибитины участвуют в процессах старения листа [Chen, Jiang, Reid, 2005]. Анализ экспрессии генов семейства прохибитинов у томата показал, что как минимум пять генов этого семейства экспрессируются в плодах, причём меняют уровни экспрессии в зависимости от стадии развития плода [Huang и др., 2021], что позволяет предположить их участие в регуляции этого процесса. Отметим, для транскрипта DN2764 в данной работе была определена значимая дифференциальная экспрессия между линией i:Bw*Alm* и сортом Bowman. Учитывая также, что у растений происходит обмен сигналами между митохондриями и пластидами [Börner и др., 2015], функциональная связь прохибитинов с такими процессами, как морфогенез пластид, синтез и накопление хлорофилла и фотосинтез, представляется интересной и перспективной темой дальнейших исследований.

Отметим также, что белковый продукт именно этого транскрипа имеет наибольшую среди всех гомологию к функционально аннотированному белку из базы данных Protein nr. Кроме того, для него наличие доменной структуры было определено с наибольшей достоверностью с помощью NCBI Structure. Эти результаты далее подтверждаются

результатами, полученными с использованием сервиса InterProScan. Этот транскрипт был выбран для локализации в геноме ячменя с помощью экспериментальных методов. Были разработаны праймеры, позволяющие определить полморфизм в строении этого гена у линии Bowman и линии i:Bw*Alm*. Полимеразная цепная реакция, проведённая на ДНК пшенично-ячменных замещённых линий, показала, что этот ген локализован в коротком плече хромосомы 3H ячменя. Поскольку известно, что геном почти изогенной линии i:Bw*Alm* отличается от генома сорта Bowman участком короткого плеча хромосомы 3H [Druka и др., 2011], можно предположить, что этого ген локализован именно в районе Alm.

Таким образом, можно предположить, что гены, кодирующие транскрипты DN4161 и, в большей степени, DN2764 – кандидаты на роль гена *Alm*. Однако, данное предположение требует дальнейшей проверки с помощью экспериментальных методов – выделения данных генов у ячменя, создания генетических конструкций, содержащих данные гены с последующей трансформацией бактерий, синтеза белков, кодируемых этими генами, в культуре трансформированных бактерий, выделения этих белков и установления их структуры с помощью спектроскопии или иных методов, поиск этих генов и их гомологов в геномах различных сортов ячменя и других злаков, хранящихся в биоресурсных коллециях, сайт-направленный мутагенез с целью получения растений ячменя, мутантных по данным генам, и изучения фенотипического проявления этих мутаций, а также прочих экспериментальных подходов. Всё это, однако, выходит за рамки данной работы.

4.3 Транскриптомный анализ линии i:Bw*Blp*

4.3.1. Функциональный анализ ДЭГ

Сравнительный анализ транскриптомов почти изогенной линии i:Bw*Blp* и сорта Bowman показал, что наиболее оптимальным конвейером обработки данных библиотек RNA-seq является конвейер, состоящий из фильтрации библиотек с помощью программы Prinseq, удаления последовательностей рРНК путём картирования библиотек на последовательности некодирующих РНК ячменя, картирования библиотек на референсную последовательность генома ячменя с помощью программы Hisat2 и поиска дифференциальной экспрессии с помощью пакета EdgeR для языка R. В ходе анализа

транскриптома были выявлены 1274 ДЭГ, из которых 849 повышают экспрессию у линии i:Bw*Blp*, и 425 – понижают.

Для списка генов с повышенной в линии i:Bw*Blp* экспрессией не было обнаружено достоверно обогащённых метаболических путей. Однако, было обнаружено участие этих генов в таких метаболических путях, как биосинтез флавонолов, флавоноидов и фенилпропаноидов. Многие фенольные соединения являются антиоксидантами [Rice-Evans, Miller, Paganga, 1997]. При этом, для линии черноколосых ячменей отмечается повышенная устойчивость к оксидативным стрессам [Ceccarelli, Grando, Van Leur, 1987]. Кроме того, некоторые флавоноиды, в частности кемпферол, повышают активность полифенолоксидаз у грибов [Lu и др., 2021], растений [Esmaeili, Ebrahimzadeh, Abdi, 2017] и животных [Tang и др., 2021]. Полифенолоксидазы – ферменты, непосредственно участвующие в синтезе меланинов [Boeckx и др., 2017b]. Путь синтеза дигидрокампферола и входящие в него ДЭГ показы на рисунке 22.

Насколько авторам известно, до сих пор не было показано связи именно кемпферола с регуляцией синтеза меланинов в растениях. Кемпферол также имеет другой функционал, в том числе, является предшественником в синтезе различных флавонолов [Berger и др., 2022]. Тем не менее, повышение экспрессии генов, участвующих в пути синтеза дигидрокемпферола, прямого предшественника кемпферола, в лемме ячменя линии с частичным меланизмом колоса, позволяет высказать предположение о положительном влиянии кемпферола на меланогенез. Для проверки этого предположения необходимы дальнейшие экспериментальные процедуры.

Гены, повышающие экспрессию в линии i:Bw*Blp* по сравнению с сортом Bowman, участвуют также в путях биосинтеза *o*-дихинонов и мономеров суберина. Хиноны, в том числе *o*-дихиноны, являются одними из предшественников в реакциях синтеза алломеланинов [Rouet-Mayer, Ralambosoa, Philippon, 1990]. Образование суберина параллельно с меланогенезом при повреждении тканей наблюдается у сахарной свёклы [Fugate и др., 2016]. Значимость этих процессов в целостных, не повреждённых тканях на данный момент изучена недостаточно. Более того, неизвестны конкретные гены ячменя, участвующие в некоторых стадиях приведённых процессов – так, в базе данных PlantCyc в метаболическом пути биосинтеза мономеров суберина приведено в общей сложности 20 реакций, из которых только для 8 приведены гены, кодирующие ферменты,

катализирующие эти реакции. Недостаточная аннотация этого и многих других метаболических путей затрудняет осмысление полученных результатов.

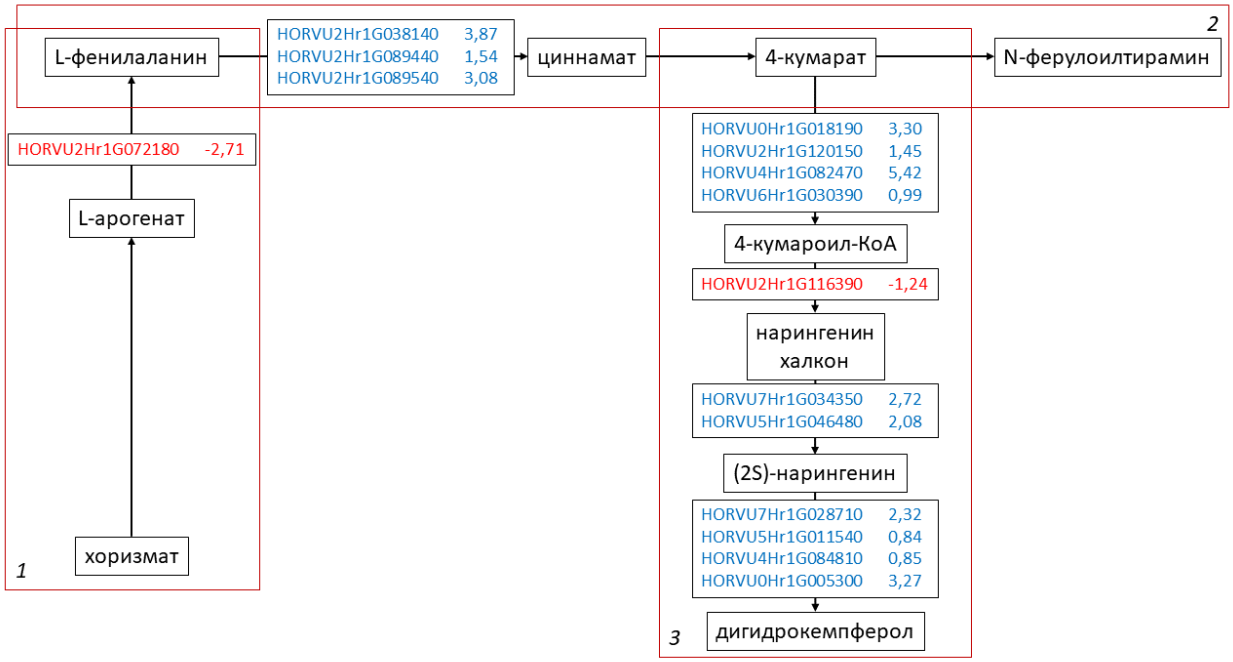

Рис. 22. Метаболический путь синтеза дигидрокемпферола из хоризмата [Rommens и др., 2008]. Приведены следующие метаболические пути, аннотированные в базе данных PlantCyc: (1) – путь синтеза фенилаланина; (2) – путь синтеза мономеров суберинов; (3) – путь синтеза флавоноидов. Приведены ДЭГ, участвующие в отдельных стадиях метаболического пути. Гены, значимо повышающие экспрессию в линии i:Bw*Blp* по сравнению с сортом Bowman, выделены синим цветом; гены, понижающие экспрессию в этой линии, выделены красным.

В целом же, многообразие терминов генной онтологии и метаболических путей, с которыми связаны гены с повышенной в линии i:Bw*Blp* экспрессией, может указывать на плейотропный эффект гена *Blp*, что согласуется с имеющимися в литературе данными [Bishaw, Struik, Gastel van, 2014; Choo, 2010].

Для генов, имеющих пониженную экспрессию у линии i:Bw*Blp*, показано участие в цикле Кальвина-Бенсона, шунте РБФК и цикле усвоения аммония. Кроме того, экспрессия генов пластома понижается у линии i:Bw*Blp*, причём гены, связанные с фотосинетзом,

понижают свою экспрессию сильнее, чем гены, кодирующие рибосомные белки. Всё это в совокупности может говорить о понижении фотосинтетической активности и в целом физиологической активности пластид в клетках цветковой чешуи и перикарпа ячменя линии i:Bw*Blp*. Это предположение подкрепляется наблюдаемыми изменениями уровней экспрессии генов, локализованных в пластоме.

Данный результат, указывающий на перестройку транскриптома пластид по сравнению с активными фотосинтезирующими хлоропластами, согласуется с полученными раннее с помощью электроной микроскопии результатами [Shoeva и др., 2020], показывающими накопление меланина в пластидах и структурные изменения пластид в клетках леммы ячменя линии i:Bw*Blp*. Выдвигается предположение о возможности выделения нового класса пластид, функция которых заключается в синтезе и накоплении меланинов – «меланопластов» [Shoeva и др., 2020]. В последние годы были проведены исследования, посвящённые растительным меланинам на примере разнообразных растений [Pandey, Dhakal, 2001; Varga и др., 2016; Wang, Rhim, 2019; Yao, Qi, 2016]. Однако, эти исследования сфокусированы в основном на структуре и химических свойствах этих соединений. Дальнейшие работы в этом направлении, проведённые на различных растительных объектах, могли бы пролить свет на участие пластид в этих процессах, и связанных с этим морфологических изменениях, происходящих в пластидах.

Рексеквенирование геномов сортов ячменя, контрастных по окраске колоса, показало, что ген HORVU1Hr1G087010 содержит 25 полиморфных SNP – больше, чем любой другой ген, локализованный в районе Blp [Long и др., 2019]. На основании этого Long и соавторы. [2019] выдвинули предположение, что ген HORVU1Hr1G087010 является геном *Blp*, обуславливающим формирование меланиновой чёрной окраски цветковой чешуи и перикарпа ячменя [Long и др., 2019]. Этот ген кодирует фосфатазу пурпурной кислоты. Наблюдаемое в данной работе повышение экспрессии этого гена у линии i:Bw*Blp* свидетельствует, что ген HORVU1Hr1G087010 может быть ассоциирован с синтезом меланинов, хотя функциональная связь фосфатазы пурпурной кислоты и синтеза меланинов остаётся неясной, и данных для подтверждения гипотезы Long et al. [2019] недостаточно.

Фосфатазы пурпурной кислоты – ферменты, широко представленные как у животных, бактерий и грибов, так и растений [Dai и др., 2011]. Считается, что их функции в первую очередь состоят в усвоении фосфора корнями растений из почвы, и их экспрессия

повышается в ответ на дефицит фосфора [Wang и др., 2011b]. Кроме того, наблюдается антиоксидантное действие фосфатаз пурпурной кислоты арабидопсис и томата, и высказывается предположение, что эти ферменты могут также участвовать в развитии растения в нормальных условиях [Zhu и др., 2005]. Анализ экспрессии 28 генов арабидопсиса, кодирующих фосфатазы пурпурной кислоты, показал, что, несмотря на то, что эти гены различаются своими паттернами экспрессии, все они экспрессируются в цветке растения [Zhu и др., 2005].

Далее, было показано наличие в последовательности гена HORVU1Hr1G086780 полиморфизма, ассоциированного с меланиновой окраской оболочек зерновки с достоверностью $p = 6{,}75 \cdot 10^{-150}$ [Глаголева, 2022]. Данная ассоциация более достоверна, чем ассоциация полиморфизмов в других генах, локализованных в локусе Blp [Глаголева, 2022]. Этот ген кодирует белок из семейства CLAVATA/CLE-пептидов, которые участвуют в формировании соцветий и опредении архитектуры колоса у злаков [Глаголева, 2022]. Таким образом, этот ген является перспективным кандидатом на роль гена *Blp*. В настоящей работе, однако, наблюдается нулевая экспрессия гена HORVU1Hr1G086780 в лемме и перикарпе как ячменя линии i:Bw*Blp*, так и растений сорта Bowman. Таким образом, имеющихся данных недостаточно, чтобы подтвердить или опровергнуть гипотезу, что ген HORVU1Hr1G086780 является геном *Blp*.

4.3.2 Анализ *de novo* реконструированного транскриптома.

Была проведена реконструкция транскриптома *de novo*, для чего были использованы несколько программ-сборщиков, после чего из полученных сборок была составлена общая мета-сборка транскриптома. По совокупности параметров, отражающих качество и полноту сборки, мета-сборка опережает индивидуальные сборки, из которых она была составлена.

В транскриптоме были обнаружены четыре новых контига, имеющие достоверную дифференциальную экспрессию, аминокислотные продукты которых имеют значимую гомологию с последовательностями белков из базы данных NCBI Protein. Рассмотрим эти контиги подробнее. Пептидный продукт контига DN12020 имеет гомологию к функционально аннотированному белку растений с крайне низким уровнем достоверности – E = 4,74. В этом пептидном продукте не было обнаружено доменной структуры. Его

кодирующий потенциал очень низок. Всё это позволяет отбросить этот контиг как артефакт сборки транскриптома.

В аминокислотном продукте контига DN16564 была обнаружена доменная структура серин/треониновых фосфатаз, и он имеет гомологию к предположительной фосфатазе белков 2C ячменя. Поэтому, можно утверждать, что этот контиг с высокой долей вероятности действительно кодирует серин-треониновую фосфатазу белков. Аналогичным образом, продукт контига DN19100 можно считать серин/треониновой протеинкиназой из-за идентифицированной доменной структуры белков этого семейства и гомологии с белков с протеинкиназой ячменя. Оба этих контига имеют значимо повышают экспрессию в линии i:Bw*Blp* в 2,36 раз и в 7 раз, соответственно. Однако, их кодирующий потенциал оценён, как очень низкий, но эта оценка основана на аминокислотных продуктах, имеющих значительно меньший размер, чем те, которые были обнаружены для этих контигов с помощью конвейера программ EvidentialGene.

Как уже было сказано ранее, функции протеинкиназ разнообразны и заключаются, в том числе, в ответе на стресс [Guo, Liu, Chong, 2018; Hake, Romeis, 2019], регуляции клеточного цикла и других процессов. Протеинфосфатазы катализируют отделение фосфатной группы от белков, фосфорилированных протеинкиназами. Они также регулируют ответ на различные виды стресса [Reyes и др., 2006]. Кроме того, они участвуют в формировании реакции на фитогормоны, в особенности на гибереллиновую кислоту и абсцизовую кислоту [Qiu и др., 2022]. В целом, функции этих белков очень многообразны, и в данном случае сложно утверждать о конкретной роли этого белка в формировании наблюдаемого фенотипа ячменя. Обращает на себя внимание повышение экспрессии обоих этих контигов в линии i:Bw*Blp*, что позволяет предположить их возможное антагонистическое действие. Однако, для проверки этого предположения необходимы дальнейшие экспериментальные процедуры. Отметим также, что протеиновые фосфотазы класса PP2C содержат в составе функциональной формы белка хелатированный ион магния Mg^{2+} [Zhang и др., 2022]. В данной же работе наблюдается обогащение термина генной онтологии «связаывание ионов магния» генами, имеющими повышенную экспрессию в линии i:Bw*Blp*.

Наконец, аминокислотный продукт контига DN21394 имеет гомологию к цитохром P450 709B1-подобному белку *T. dicoccoides* и с высокой достоверностью содержит в своём

составе обнаружен домен цитохрома P450. Это позволяет утверждать с большой долей вероятности, что ген, транскрипт которого реконструирован в этом контиге, действительно кодирует цитохром. Цитохром участвует в цепи переноса электронов в митохондриях, в процессе аэробного дыхания [Yoshida, Terashima, Noguchi, 2006], и в пластидах в процессе фотосинтеза [Yoshida, Terashima, Noguchi, 2007]. В линии i:Bw*Blp* подавлен процесс фотосинтеза, как следует из анализа ДЭГ. Как говорилось ранее, функции пластид и митохондрий неразрывно связаны и взаимно регулируют друг друга [Hedtke и др., 2002]. Это может объяснять понижение экспрессии этого контига, кодирующего цитохромный белок, в линии i:Bw*Blp*.

Заключение

Проведённый в данной работе биоинформатический анализ библиотек коротких прочтений позволяет сделать определённые выводы как о методологической части обработки данных RNA-seq, так и об изменениях транскриптома леммы ячменя, сопутствующих формированию контрастных по окраске колоса фенотипов.

Результаты, полученные в данной работе, свидетельствуют, что использование нескольких конвейеров биоинформатической обработки данных с отбором наиболее оптимального конвейера для конкретных имеющихся данных позволяют достичь наибольшей точности анализа и определения дифференциальной экспрессии генов. Кроме того, использование нескольких программ для *de novo* реконструкции транскриптома и последующее объединение полученных *de novo* сборок даёт наибольшую полноту и точность реконструкции транскриптов из данных RNA-seq. Авторы предполагают, что такой подход применим в работе с данными массового параллельного секвенирования транскриптомов с помощью платформ для секвенирования второго поколения на самых разнообразных биологических объектах, относящихся к эукариотическим организмам. Также авторы считают, что этот подход позволит улучшить результаты анализа и приблизиться к пониманию биологического смысла и молекулярных процессов, стоящих за исследуемыми явлениями.

Функциональный анализ генов, изменяющих экспрессию в линии, характеризующейся частичным альбинизмом, показал ассиметрию изменения профилей экспрессии, происходящих в клетках леммы этой линии. Количество генов, понижающих свою экспрессию в этой линии, превышает количество генов, понижающих экспрессию, более чем в десять раз. Гены, понижающие экспрессию, связаны с фотосинтезом, метаболизмом азота и аэробным дыханием. Однако, учитывая специфику исследуемого фенотипа, можно утверждать, что все эти процессы – это следствие, а не причина формирования фенотипа. Гены же с повышенной в этой линии экспрессией участвуют в протеолизе и ответе на стресс, что также можно связать с наблюдаемым фенотипом. Анализ *de novo* реконструированного транскриптома позволил обнаружить транскрипты, не представленные в текущей версии генома ячменя. Независимая экспериментальная проверка показала, что ген с повышенной экспрессией в этой линии, кодирующий

прохибитиновый белок, локализован в коротком плече хромосомы 3H ячменя. Ген, мутация в котором приводит к формированию изучаемого фенотипа, локализован именно в этом плече хромосомы 3H. Тем не менее, без дополнительной верификации с помощью экспериментальных процедур, таких как сайт-направленный мутагенез или генетическая трансформация, нельзя утверждать, является ли этот ген искомым геном *Alm*.

Анализ транскриптома линии, характеризующейся меланизмом колоса, позволил выделить гены, имеющие достоверную дифференциальную экспрессию. Гены с пониженной экспрессией в этой линии, связаны с фотосинтезом и сопутствующими процессами – шунтом РБФК и усвоением аммония. Это говорит о подавлении фотосинтеза в органах растений, содержащих меланин. Для генов же с повышенной экспрессией в этой линии, показано участие в синтезе разнообразных метаболитов, в том числе флавоноидов и мономеров суберина. Более подробному функциональному описанию этих генов препятствует недостаточная изученность и аннотация метаболических путей ячменя, в особенности связанных с биосинтезом меланинов и их предшественников. *De novo* реконструкция транскриптома этой линии выявила транскрипты, не представленные в текущей версии генома ячменя, часть из которых имеет значимую дифференциальную экспрессию. В этой линии понижена экспрессия гена, кодирующего цитохром P450, и повышена экспрессия двух генов, один из которых кодирует протеинкиназу, другой – фосфатазу белков. Хотя для белков этих типов известно участие в регуляции множества физиологических процессов растений, делать выводы о том, является ли один из этих генов искомым геном *Blp*, нельзя без экспериментальной верификации.

Выводы

1. Разработан метод биоинформатического конвейерного анализа транскриптомных данных на основе комбинации наборов компьютерных программ для оценки уровня экспрессии генов путём 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 3H ячменя 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 *о*-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 использованной литературы

1. Arnaud O. и др. Targeted reduction of highly abundant transcripts using pseudo-random primers // BioTechniques. 2016. Т. 60. № 4. С. 169–174.
2. Wald A. Sequential Tests of Statistical Hypotheses // Ann. Math. Stat. 1945.
3. Engström P. G. и др. Systematic evaluation of spliced alignment programs for RNA-seq data // Nat. Methods. 2013.
4. Mortazavi A. и др. Mapping and quantifying mammalian transcriptomes by RNA-Seq // Nat. Methods. 2008. Т. 5. № 7. С. 621–628.
5. Dobin A. и др. STAR: Ultrafast universal RNA-seq aligner // Bioinformatics. 2013. Т. 29. № 1. С. 15–21.
6. Rapaport F. и др. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data // Genome Biol. 2013.
7. Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform // Bioinformatics. 2009. Т. 25. № 14. С. 1754–1760.
8. Love M. I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 // Genome Biol. 2014.
9. Simão F. A. и др. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs // Bioinformatics. 2015. Т. 31. № 19. С. 3210–3212.
10. Lahens N. F. и др. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression // BMC Genomics. 2017.
11. Grabherr M. G. ; и др. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data // Nat. Biotechnol. 2013. Т. 29. № 7. С. 644–652.
12. Mielczarek M., Szyda J. Review of alignment and SNP calling algorithms for next-generation sequencing data // J. Appl. Genet. 2016.
13. Lafond-Lapalme J. и др. A new method for decontamination of de novo transcriptomes using a hierarchical clustering algorithm // Bioinformatics. 2017.
14. Lin H. N., Hsu W. L. Kart: A divide-and-conquer algorithm for NGS read alignment // Bioinformatics. 2017.
15. Fu S. и др. IDP-denovo: De novo transcriptome assembly and isoform annotation by hybrid sequencing // Bioinformatics. , 2018.
16. Lin H. N., Hsu W. L. DART: A fast and accurate RNA-seq mapper with a partitioning strategy // Bioinformatics. 2018a. Т. 34. № 2. С. 190–197.
17. Li H., Homer N. A survey of sequence alignment algorithms for next-generation sequencing // Brief. Bioinform. 2010.
18. Edgar R. C. Search and clustering orders of magnitude faster than BLAST // Bioinformatics. 2010.
19. Smith-Unna R. и др. TransRate: Reference-free quality assessment of de novo transcriptome assemblies // Genome Res. 2016.
20. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads // EMBnet.journal. 2011.

21. Bullard J. H. и др. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments // BMC Bioinformatics. 2010.
22. Auer P. L., Doerge R. W. Statistical Design and Analysis of RNA Sequencing Data // Genetics. 2010.
23. Mascher M. и др. A chromosome conformation capture ordered sequence of the barley genome // Nature. 2017.
24. Robinson M. D., Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data // Genome Biol. 2010.
25. Langmead B. и др. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome // Genome Biol. 2009. Т. 10. № 3. С. R25.
26. Вавилов Н. И. Центры происхождения культурных растений. Ленинград: тип. им. Гутенберга, 1926. 248 с.
27. Ahmad S. и др. Tryptophan, a non-canonical melanin precursor: New L-tryptophan based melanin production by Rubrivivax benzoatilyticus JA2 // Sci. Rep. 2020.
28. Aldridge S., Hadfield J. Introduction to miRNA profiling technologies and cross-platform comparison // Methods Mol. Biol. 2012.
29. Allorent G. и др. Plastid gene expression during chloroplast differentiation and dedifferentiation into non-photosynthetic plastids during seed formation // Plant Mol. Biol. 2013. Т. 82. С. 59–70.
30. Altschul S. F. и др. Basic local alignment search tool // J. Mol. Biol. 1990. Т. 215. № 3. С. 403–410.
31. Amaral A. J. и др. Quality assessment and control of tissue specific RNA-seq libraries of drosophila transgenic RNAi models // Front. Genet. 2014. Т. 5. № MAR. С. 1–12.
32. Amiryousefi A., Hyvönen J., Poczai P. The chloroplast genome sequence of bittersweet (Solanum dulcamara): Plastid genome structure evolution in Solanaceae // PloS One. 2018. Т. 13. № 4. С. e0196069.
33. Anders S., Pyl P. T., Huber W. HTSeq-A Python framework to work with high-throughput sequencing data // Bioinformatics. 2015.
34. Archer S. K., Shirokikh N. E., Preiss T. Probe-directed degradation (PDD) for flexible removal of unwanted cDNA sequences from RNA-seq libraries // Curr. Protoc. Hum. Genet. 2015. Т. 2015. С. 11.15.1-11.15.36.
35. Armero A. и др. Improving transcriptome de novo assembly by using a reference genome of a related species: Translational genomics from oil palm to coconut // PLoS ONE. 2017.
36. Artal-Sanz M., Tavernarakis N. Prohibitin and mitochondrial biology // Trends Endocrinol. Metab. 2009. Т. 20. № 8. С. 394–401.
37. Badr A. и др. On the Origin and Domestication History of Barley (Hordeum vulgare) // Mol. Biol. Evol. 2000. Т. 17. № 4. С. 499–510.
38. Bae C. H. и др. Regulation of chloroplast gene expression is affected in ali, a novel tobacco albino mutant // Ann. Bot. 2001. Т. 88. № 4. С. 545–553.
39. Banerjee G., Singh D., Sinha A. K. Plant cell cycle regulators: Mitogen-activated protein kinase, a new regulating switch? // Plant Sci. 2020. Т. 301. С. 110660.

40. Bao A. и др. The Stable Level of Glutamine synthetase 2 Plays an Important Role in Rice Growth and in Carbon-Nitrogen Metabolic Balance // Int. J. Mol. Sci. 2015. Т. 16. № 6. С. 12713–12736.
41. Baslam M. и др. Recent Advances in Carbon and Nitrogen Metabolism in C3 Plants // Int. J. Mol. Sci. 2020. Т. 22. № 1. С. 318.
42. Baune M.-C. и др. The Arabidopsis Plastidial Glucose-6-Phosphate Transporter GPT1 is Dually Targeted to Peroxisomes via the Endoplasmic Reticulum[OPEN] // Plant Cell. 2020. Т. 32. № 5. С. 1703–1726.
43. Bauwe H., Hagemann M., Fernie A. R. Photorespiration: players, partners and origin // Trends Plant Sci. 2010. Т. 15. № 6. С. 330–336.
44. Benjamin A. M. и др. Comparing reference-based RNA-Seq mapping methods for non-human primate data // BMC Genomics. 2014.
45. Berger A. и др. Kaempferol as a precursor for ubiquinone (coenzyme Q) biosynthesis: An atypical node between specialized metabolism and primary metabolism // Curr. Opin. Plant Biol. 2022. Т. 66. С. 102165.
46. Berry J. O., Mure C. M., Yerramsetty P. Regulation of Rubisco gene expression in C4 plants // Curr. Opin. Plant Biol. 2016. Т. 31. С. 23–28.
47. Betti M. и др. Manipulating photorespiration to increase plant productivity: recent advances and perspectives for crop improvement // J. Exp. Bot. 2016. Т. 67. № 10. С. 2977–2988.
48. Bian J. и др. Transcriptional dynamics of grain development in barley (Hordeum vulgare L.) // Int. J. Mol. Sci. 2019.
49. Bishaw Z., Struik P. C., Gastel A. J. G. van. Wheat and barley seed system in Syria: How diverse are wheat and barley varieties and landraces from farmer's fields? // Int. J. Plant Prod. 2014.
50. Bleidorn C. Third generation sequencing: Technology and its potential impact on evolutionary biodiversity research // Syst. Biodivers. 2016.
51. Bobik K., Burch-Smith T. M. Chloroplast signaling within, between and beyond cells // Front. Plant Sci. 2015. Т. 6. С. 781.
52. Boeckx T. и др. Polyphenol oxidase in leaves: Is there any significance to the chloroplastic localization? // J. Exp. Bot. 2015.
53. Boeckx T. и др. Detection of potential chloroplastic substrates for polyphenol oxidase suggests a role in undamaged leaves // Front. Plant Sci. 2017a.
54. Boeckx T. и др. Detection of Potential Chloroplastic Substrates for Polyphenol Oxidase Suggests a Role in Undamaged Leaves // Front. Plant Sci. 2017b. Т. 8.
55. Bolger A. M., Lohse M., Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data // Bioinformatics. 2014.
56. Börner T. и др. Chloroplast RNA polymerases: role in chloroplast biogenesis // Biochim. Biophys. Acta BBA-Bioenerg. 2015. Т. 1847. № 9. С. 761–769.
57. Börner T. The discovery of plastid-to-nucleus retrograde signaling—a personal perspective // Protoplasma. 2017. Т. 254. № 5. С. 1845–1855.
58. Bourgeois Y. X. C. и др. Candidate Gene Analysis Suggests Untapped Genetic Complexity in Melanin-Based Pigmentation in Birds // J. Hered. 2016.

59. Bradbeer J. W. и др. Cytoplasmic synthesis of plastid polypeptides may be controlled by plastid-synthesised RNA // Nature. 1979. Т. 279. № 5716. С. 816–817.
60. Braun H.-P. The oxidative phosphorylation system of the mitochondria in plants // Mitochondrion. 2020. Т. 53. С. 66–75.
61. Bryant N. и др. Identification of nuclear genes encoding chloroplast-localized proteins required for embryo development in Arabidopsis // Plant Physiol. 2011. Т. 155. № 4. С. 1678–1689.
62. Brzezowski P., Richter A. S., Grimm B. Regulation and function of tetrapyrrole biosynthesis in plants and algae // Biochim. Biophys. Acta BBA-Bioenerg. 2015. Т. 1847. № 9. С. 968–985.
63. Buckley G. F. H. Inheritance in Barley With Special Reference to the Color of Caryopsis and Lemma // Sci. Agric. 1930. Т. 10. № 7. С. 460–492.
64. Buer C. S., Imin N., Djordjevic M. A. Flavonoids: new roles for old molecules // J. Integr. Plant Biol. 2010. Т. 52. № 1. С. 98–111.
65. Bushmanova E. и др. RnaQUAST: A quality assessment tool for de novo transcriptome assemblies // Bioinformatics. 2016.
66. Bushmanova E. и др. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data // bioRxiv. 2018.
67. Butler M. J., Gardiner R. B., Day A. W. Melanin synthesis by Sclerotinia sclerotiorum // Mycologia. 2009.
68. Cackett L. и др. Chloroplast development in green plant tissues: the interplay between light, hormone, and transcriptional regulation // New Phytol. 2022. Т. 233. № 5. С. 2000–2016.
69. Camacho C. и др. BLAST+: Architecture and applications // BMC Bioinformatics. 2009. Т. 10. С. 1–9.
70. Cao P. и др. Purine nucleotide biosynthetic gene GARS controls early chloroplast development in rice (Oryza sativa L.) // Plant Cell Rep. 2019. Т. 38. С. 183–194.
71. Cao W. и др. Unraveling the Structure and Function of Melanin through Synthesis // J. Am. Chem. Soc. 2021.
72. Casanova-Sáez R. и др. Arabidopsis ANGULATA10 is required for thylakoid biogenesis and mesophyll development // J. Exp. Bot. 2014. Т. 65. № 9. С. 2391–2404.
73. Ceccarelli S., Grando S., Van Leur J. A. G. Genetic diversity in barley landraces from Syria and Jordan // Euphytica. 1987.
74. Cerveau N., Jackson D. J. Combining independent de novo assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms // BMC Bioinformatics. 2016.
75. Chan K. X. и др. Learning the Languages of the Chloroplast: Retrograde Signaling and Beyond // Annu. Rev. Plant Biol. 2016.
76. Charkoudian L. K., Franz K. J. Fe(III)-Coordination Properties of Neuromelanin Components: 5,6-Dihydroxyindole and 5,6-Dihydroxyindole-2-carboxylic Acid // Inorg. Chem. 2006. Т. 45. № 9. С. 3657–3664.
77. Chen G. и др. Tibet is one of the centers of domestication of cultivated barley // Proc. Natl. Acad. Sci. 2012a.

78. Chen J.-C., Jiang C.-Z., Reid M. S. Silencing a prohibitin alters plant development and senescence: Prohibitins in development and senescence // Plant J. 2005. Т. 44. № 1. С. 16–24.
79. Chen K. и др. NAL8 encodes a prohibitin that contributes to leaf and spikelet development by regulating mitochondria and chloroplasts stability in rice // BMC Plant Biol. 2019. Т. 19. № 1. С. 395.
80. Chen L. Y. и др. RNASEQR-a streamlined and accurate RNA-seq sequence analysis program // Nucleic Acids Res. 2012b.
81. Chen M. и др. Arabidopsis HEMERA/pTAC12 initiates photomorphogenesis by phytochromes // Cell. 2010. Т. 141. № 7. С. 1230–1240.
82. Chen S. и др. Fastp: An ultra-fast all-in-one FASTQ preprocessor // Bioinformatics. , 2018.
83. Chen T. и др. Fine mapping and candidate gene analysis of a green-revertible albino gene gra(t) in rice // J. Genet. Genomics. 2009a. Т. 36. № 2. С. 117–123.
84. Chen X. и др. Identification and characterization of novel amphioxus microRNAs by Solexa sequencing // Genome Biol. 2009b. Т. 10. № 7. С. R78.
85. Chen X. и др. Protein kinases in plant responses to drought, salt, and cold stress // J. Integr. Plant Biol. 2021. Т. 63. № 1. С. 53–78.
86. Chen Z. и др. Genetic diversity analysis of barley landraces and cultivars in the Shanghai region of China // Genet. Mol. Res. GMR. 2012c. Т. 11. С. 644–50.
87. Cheung F. и др. Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology // BMC Genomics. 2006. Т. 7. С. 1–10.
88. Chi W., Sun X., Zhang L. Intracellular signaling from plastid to nucleus // Annu. Rev. Plant Biol. 2013. Т. 64. С. 559–582.
89. Choo T. M. Breeding Barley for Resistance to Fusarium Head Blight and Mycotoxin Accumulation // Plant Breeding Reviews. , 2010.
90. Cloonan N. и др. Stem cell transcriptome profiling via massive-scale mRNA sequencing // Nat. Methods. 2008. Т. 5. № 7. С. 613–619.
91. Conesa A. и др. A survey of best practices for RNA-seq data analysis // Genome Biol. 2016. Т. 17. № 1. С. 1–19.
92. Cui J. и др. Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome // Plant Methods. 2020.
93. Dai F. и др. Identification of a Phytase gene in barley (Hordeum vulgare L.) // PLoS ONE. 2011.
94. D’Alba L., Shawkey M. D. Melanosomes: Biogenesis, properties, and evolution of an ancient organelle // Physiol. Rev. 2019.
95. Darling A. C. E. и др. Mauve: Multiple alignment of conserved genomic sequence with rearrangements // Genome Res. 2004.
96. Dawson I. K. и др. Barley: a translational model for adaptation to climate change // New Phytol. 2015. Т. 206. № 3. С. 913–931.
97. Del Fabbro C. и др. An extensive evaluation of read trimming effects on illumina NGS data analysis // PLoS ONE. 2013.

98. Demarsy E. и др. Building up of the plastid transcriptional machinery during germination and early plant development // Plant Physiol. 2006. Т. 142. № 3. С. 993–1003.
99. Demarsy E. и др. Characterization of the plastid-specific germination and seedling establishment transcriptional programme // J. Exp. Bot. 2012. Т. 63. № 2. С. 925–939.
100. Dillies M. A. и др. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis // Brief. Bioinform. 2013. Т. 14. № 6. С. 671–683.
101. D'Mello S. A. N. и др. Signaling pathways in melanogenesis // Int. J. Mol. Sci. 2016.
102. Dong Z. C., Chen Y. Transcriptomics: Advances and approaches // Sci. China Life Sci. 2013.
103. Drewe P. и др. Accurate detection of differential RNA processing // Nucleic Acids Res. 2013. Т. 41. № 10. С. 5189–5198.
104. Druka A. и др. Genetic Dissection of Barley Morphology and Development // PLANT Physiol. 2011.
105. Dubos C. и др. MYB transcription factors in Arabidopsis // Trends Plant Sci. 2010. Т. 15. № 10. С. 573–581.
106. Dubreuil C. и др. Establishment of photosynthesis through chloroplast development is controlled by two distinct regulatory phases // Plant Physiol. 2018.
107. Dunwell J. M. и др. Evolution of functional diversity in the cupin superfamily // Trends Biochem. Sci. 2001. Т. 26. № 12. С. 740–746.
108. Dunwell J. M., Purvis A., Khuri S. Cupins: the most functionally diverse protein superfamily? // Phytochemistry. 2004. Т. 65. № 1. С. 7–17.
109. Dyson B. C., Webster R. E., Johnson G. N. GPT2: a glucose 6-phosphate/phosphate translocator with a novel role in the regulation of sugar signalling during seedling development // Ann. Bot. 2014. Т. 113. № 4. С. 643–652.
110. Eisenman H. C., Casadevall A. Synthesis and assembly of fungal melanin // Appl. Microbiol. Biotechnol. 2012.
111. Ellis R. J. Protein synthesis by isolated chloroplasts // Biochim. Biophys. Acta BBA-Rev. Bioenerg. 1977. Т. 463. № 2. С. 185–215.
112. Emanuel C. и др. Chloroplast development affects expression of phage-type RNA polymerases in barley leaves // Plant J. 2004. Т. 38. № 3. С. 460–472.
113. Ems S. C. и др. Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant Epifagus virginiana // Plant Mol. Biol. 1995. Т. 29. С. 721–733.
114. Eskelin K. и др. Ribosome profiles and riboproteomes of healthy and Potato virus A- and Agrobacterium-infected Nicotiana benthamiana plants // Mol. Plant Pathol. 2019. Т. 20. № 3. С. 392–409.
115. Esmaeili N., Ebrahimzadeh H., Abdi K. Correlation between polyphenol oxidase (PPO) activity and total phenolic contents in Crocus sativus L. corms during dormancy and sprouting stages // Pharmacogn. Mag. 2017. Т. 13. № 51. С. 519.
116. Esnaola M. и др. A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments // BMC Bioinformatics. 2013.

117. Evangelistella C. и др. 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 // Biotechnol. Biofuels. 2017.
118. Ewing B., Green P. Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities // Genome Res. 1998. Т. 8. № 3. С. 186–194.
119. Fabiańska I., Bucher M., Häusler R. E. Intracellular phosphate homeostasis – A short way from metabolism to signaling // Plant Sci. 2019. Т. 286. С. 57–67.
120. Facchinelli F., Weber A. P. M. The Metabolite Transporters of the Plastid Envelope: An Update // Front. Plant Sci. 2011. Т. 2.
121. Farrer R. A. и др. De novo assembly of the Pseudomonas syringae pv. syringae B728a genome using Illumina/Solexa short sequence reads: RESEARCH LETTER // FEMS Microbiol. Lett. 2009. Т. 291. № 1. С. 103–111.
122. Fickett J. W. Recognition of protein coding regions in DNA sequences // Nucleic Acids Res. 1982. Т. 10. № 17. С. 5303–5318.
123. Flügge U.-I. и др. Functional genomics of phosphate antiport systems of plastids // Physiol. Plant. 2003. Т. 118. № 4. С. 475–482.
124. Flügge U.-I. и др. The role of transporters in supplying energy to plant plastids // J. Exp. Bot. 2011. Т. 62. № 7. С. 2381–2392.
125. Fox S., Filichkin S., Mockler T. C. Applications of ultra-high-throughput sequencing. // Methods Mol. Biol. Clifton NJ. 2009. Т. 553. С. 79–108.
126. Franckowiak J. D., Lundqvist U., Kleinhofs A. Barley Genetics Newsletter.
127. Frazee A. C. и др. Polyester: Simulating RNA-seq datasets with differential transcript expression // Bioinformatics. 2015. Т. 31. № 17. С. 2778–2784.
128. Fu L. и др. CD-HIT: Accelerated for clustering the next-generation sequencing data // Bioinformatics. 2012.
129. Fugate K. K. и др. Cold Temperature Delays Wound Healing in Postharvest Sugarbeet Roots // Front. Plant Sci. 2016. Т. 7.
130. Gadjieva R. и др. Analysis of gun phenotype in barley magnesium chelatase and Mg-protoporphyrin IX monomethyl ester cyclase mutants // Plant Physiol. Biochem. 2005. Т. 43. № 10–11. С. 901–908.
131. Galván I., Solano F. Bird integumentary melanins: Biosynthesis, forms, function and evolution // Int. J. Mol. Sci. 2016.
132. Gang H. и др. Loss of GLK1 transcription factor function reveals new insights in chlorophyll biosynthesis and chloroplast development // J. Exp. Bot. 2019. Т. 70. № 12. С. 3125–3138.
133. Garnik E. Y. и др. Genome uncoupled (gun) phenotype is associated with root growth repression in Arabidopsis seedlings grown on lincomycin // Theor. Exp. Plant Physiol. 2019. Т. 31. С. 445–454.
134. Ghitti E. и др. Flavonoids Are Intra- and Inter-Kingdom Modulator Signals // Microorganisms. 2022. Т. 10. № 12.
135. Gilbert D. G. Genes of the pig, Sus scrofa, reconstructed with EvidentialGene // PeerJ. 2019.

136. Glagoleva A. Y., Shoeva O. Y., Khlestkina E. K. Melanin Pigment in Plants: Current Knowledge and Future Perspectives // Front. Plant Sci. 2020. Т. 11. С. 770.
137. González Carretero L., Wollstonecroft M., Fuller D. Q. A methodological approach to the study of archaeological cereal meals: a case study at Çatalhöyük East (Turkey) // Veg. Hist. Archaeobotany. 2017.
138. Gough J. и др. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure11Edited by G. Von Heijne // J. Mol. Biol. 2001. Т. 313. № 4. С. 903–919.
139. Grant G. R. и др. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM) // Bioinformatics. 2011.
140. Griebel T. и др. Modelling and simulating generic RNA-Seq experiments with the flux simulator // Nucleic Acids Res. 2012. Т. 40. № 20. С. 10073–10083.
141. Grübler B. и др. Light and plastid signals regulate different sets of genes in the albino mutant pap7-1 // Plant Physiol. 2017. Т. 175. № 3. С. 1203–1219.
142. Guan Q. и др. Heat stress induction of *miR398* triggers a regulatory loop that is critical for thermotolerance in Arabidopsis // Plant J. 2013. Т. 74. № 5. С. 840–851.
143. Guo B. и др. Comparative Proteomic Analysis of Two Barley Cultivars (Hordeum vulgare L.) with Contrasting Grain Protein Content // Front. Plant Sci. 2016. Т. 7.
144. Guo X., Liu D., Chong K. Cold signaling in plants: Insights into mechanisms and regulation: Cold stress signaling // J. Integr. Plant Biol. 2018. Т. 60. № 9. С. 745–756.
145. Gupta V. и др. RNA-Seq analysis and annotation of a draft blueberry genome assembly identifies candidate genes involved in fruit ripening, biosynthesis of bioactive compounds, and stage-specific alternative splicing // GigaScience. 2015. Т. 4. № 1. С. 1–22.
146. Haferkamp I., Fernie A. R., Neuhaus H. E. Adenine nucleotide transport in plants: much more than a mitochondrial issue // Trends Plant Sci. 2011. Т. 16. № 9. С. 507–515.
147. Haft D. H. и др. TIGRFAMs and Genome Properties in 2013 // Nucleic Acids Res. 2013. Т. 41. № D1. С. D387–D395.
148. Hagemann M., Bauwe H. Photorespiration and the potential to improve photosynthesis // Energy Mech. Biol. 2016. Т. 35. С. 109–116.
149. Hagemann R., Scholz F. A case of gene induced mutations of the plasmotype in barley // Theor Appl Genet. 1962. Т. 32. С. 50–59.
150. Hajdukiewicz P. T. J., Allison L. A., Maliga P. The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids // EMBO J. 1997. Т. 16. № 13. С. 4041–4048.
151. Hake K., Romeis T. Protein kinase-mediated signalling in priming: Immune signal initiation, propagation, and establishment of long-term pathogen resistance in plants: Kinase-mediated signaling in defense priming // Plant Cell Environ. 2019. Т. 42. № 3. С. 904–917.
152. Hart G. E., Islam A., Shepherd K. W. Use of isozymes as chromosome markers in the isolation and characterization of wheat-barley chromosome addition lines // Genet. Res. 1980. Т. 36. № 3. С. 311–325.

153. Hedtke B. и др. Six active phage-type RNA polymerase genes in Nicotiana tabacum // Plant J. 2002. Т. 30. № 6. С. 625–637.
154. Hedtke B., Börner T., Weihe A. One RNA polymerase serving two genomes // EMBO Rep. 2000. Т. 1. № 5. С. 435–440.
155. Hernández-Romero D., Solano F., Sanchez-Amat A. Polyphenol oxidase activity expression in Ralstonia solanacearum // Appl. Environ. Microbiol. 2005.
156. Hernández-Verdeja T., Strand Å. Retrograde signals navigate the path to chloroplast development // Plant Physiol. 2018. Т. 176. № 2. С. 967–976.
157. Hess W. R. и др. Inefficient rpl2 splicing in barley mutants with ribosome-deficient plastids. // Plant Cell. 1994. Т. 6. № 10. С. 1455–1465.
158. Hölzer M., Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers // GigaScience. 2019.
159. Homer N., Merriman B., Nelson S. F. BFAST: An alignment tool for large scale genome resequencing // PLoS ONE. 2009.
160. Honaas L. A. и др. Selecting superior de novo transcriptome assemblies: Lessons learned by leveraging the best plant genome // PLoS ONE. 2016.
161. Horvath S. E. и др. Role of membrane contact sites in protein import into mitochondria // Protein Sci. 2015. Т. 24. № 3. С. 277–297.
162. Hou R. и др. Impact of the next-generation sequencing data depth on various biological result inferences // Sci. China Life Sci. 2013. Т. 56. № 2. С. 104–109.
163. Hua W. и др. Identification and fine mapping of a white husk gene in barley (Hordeum vulgare L.) // PLoS ONE. 2016.
164. Huang F. и др. The prohibitins (PHB) gene family in tomato: Bioinformatic identification and expression analysis under abiotic and phytohormone stresses // GM Crops Food. 2021. Т. 12. № 1. С. 535–550.
165. Huang H.-C., Niu Y., Qin L.-X. Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software: Supplementary Issue: Sequencing Platform Modeling and Analysis // Cancer Inform. 2015. Т. 14s1. С. CIN.S21631.
166. Huang R., Yang C., Zhang S. The *Arabidopsis* PHB3 is a pleiotropic regulator for plant development // Plant Signal. Behav. 2019. Т. 14. № 11. С. 1656036.
167. Huang S., Millar A. H. Succinate dehydrogenase: the complex roles of a simple enzyme // Curr. Opin. Plant Biol. 2013. Т. 16. № 3. С. 344–349.
168. Hummel M. и др. Proteomic LC–MS analysis of Arabidopsis cytosolic ribosomes: Identification of ribosomal protein paralogs and re-annotation of the ribosomal protein genes // J. Proteomics. 2015. Т. 128. С. 436–449.
169. Ingolia N. T. и др. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling // Science. 2009.
170. Itou T., Ito S., Wakamatsu K. Effects of aging on hair color, melanosome morphology, and melanin composition in Japanese females // Int. J. Mol. Sci. 2019.
171. Iwaya T. и др. Contrasting Expression Patterns of Histone mRNA and microRNA 760 in Patients with Gastric Cancer // Clin. Cancer Res. 2013. Т. 19. № 23. С. 6438–6449.

172. Iyer M. K., Chinnaiyan A. M., Maher C. A. ChimeraScan: A tool for identifying chimeric transcription in sequencing data // Bioinformatics. 2011.
173. Jehl F. и др. RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species // Front. Genet. 2021. Т. 12. С. 655707.
174. Jiang D. и др. The evolution and functional divergence of the histone H2B family in plants // PLoS Genet. 2020. Т. 16. № 7. С. e1008964.
175. Johnson D. S., Mortazavi A., Myers R. M. Protein-DNA Interactions // 2007. № June. С. 1497–1503.
176. Jones P. и др. InterProScan 5: genome-scale protein function classification // Bioinformatics. 2014. Т. 30. № 9. С. 1236–1240.
177. José Ripoll J. и др. PEPPER, a novel K-homology domain gene, regulates vegetative and gynoecium development in Arabidopsis // Dev. Biol. 2006. Т. 289. № 2. С. 346–359.
178. Joyard J. и др. Chloroplast Proteomics and the Compartmentation of Plastidial Isoprenoid Biosynthetic Pathways // Mol. Plant. 2009. Т. 2. № 6. С. 1154–1180.
179. Kamei H. и др. Suppression of Growth of Cultured Malignant Cells by Allomelanins, Plant-produced Melanins // Cancer Biother. Radiopharm. 1997. Т. 12. № 1. С. 47–49.
180. Kang Y.-J. и др. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features // Nucleic Acids Res. 2017. Т. 45. № W1. С. W12–W16.
181. Kanitz A. и др. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data // Genome Biol. 2015.
182. Katiyar A. и др. Genome-wide classification and expression analysis of MYB transcription factor families in rice and Arabidopsis // BMC Genomics. 2012. Т. 13. № 1. С. 544.
183. Katz Y. и др. Analysis and design of RNA sequencing experiments for identifying isoform regulation // Nat. Methods. 2010. Т. 7. № 12. С. 1009–1015.
184. Kawashima T. и др. Diversification of histone H2A variants during plant evolution // Trends Plant Sci. 2015. Т. 20. № 7. С. 419–425.
185. Kelley S. K. и др. Identification of a tyrosinasse from a periphytic marine bacterium // FEMS Microbiol. Lett. 1990.
186. Kent W. J. BLAT - The BLAST-like alignment tool // Genome Res. 2002.
187. Kersey P. J. и др. Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species // Nucleic Acids Res. 2018.
188. KEYS A. J. и др. Photorespiratory nitrogen cycle // Nature. 1978. Т. 275. № 5682. С. 741–743.
189. Khan N. Z., Lindquist E., Aronsson H. New putative chloroplast vesicle transport components and cargo proteins revealed using a bioinformatics approach: an Arabidopsis model // PloS One. 2013. Т. 8. № 4. С. e59898.
190. Kim D. и др. TopHat2 : accurate alignment of transcriptomes in the presence of insertions , deletions and gene fusions // 2013a. С. 0–9.
191. Kim D., Langmead B., Salzberg S. L. HISAT: A fast spliced aligner with low memory requirements // Nat. Methods. 2015.

192. Kim S. и др. Arabidopsis Chlorophyll Biosynthesis: An Essential Balance between the Methylerythritol Phosphate and Tetrapyrrole Pathways // Plant Cell. 2013b. Т. 25. № 12. С. 4984–4993.
193. Kim Y. J., Uyama H. Tyrosinase inhibitors from natural and synthetic sources: Structure, inhibition mechanism and perspective for the future // Cell. Mol. Life Sci. 2005.
194. Klepikova A. V. и др. Effect of method of deduplication on estimation of differential gene expression using RNA-seq // PeerJ. 2017. Т. 5. С. e3091.
195. Korčeková D. и др. Nucleologenesis in the Caenorhabditis elegans embryo // PloS One. 2012. Т. 7. № 7. С. e40290–e40290.
196. Koussevitzky S. и др. An Arabidopsis thaliana virescent mutant reveals a role for ClpR1 in plastid development // Plant Mol. Biol. 2007. Т. 63. С. 85–96.
197. Krupinska K. и др. Genome communication in plants mediated by organelle–n-ucleus-located proteins // Philos. Trans. R. Soc. B Biol. Sci. 2020. Т. 375. № 1801. С. 20190397.
198. Kumar S. и др. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots // Front. Genet. 2013.
199. Kurtz S. и др. Versatile and open software for comparing large genomes. // Genome Biol. 2004.
200. Kvam V. M., Liu P., Yaqing S. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data // Am. J. Bot. 2012.
201. Lamarre S. и др. Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size // Front. Plant Sci. 2018. Т. 9. № February.
202. Landau A. и др. Two infA gene mutations independently originated from a mutator genotype in barley // J. Hered. 2007. Т. 98. № 3. С. 272–276.
203. Langaee T., Ronaghi M. Genetic variation analyses by Pyrosequencing // Mutat. Res. Mol. Mech. Mutagen. 2005. Т. 573. № 1–2. С. 96–102.
204. Langmead B., Salzberg S. L. Fast gapped-read alignment with Bowtie 2. // Nat. Methods. 2012.
205. Larkin J. C., Brown M. L., Schiefelbein J. How do cells know what they want to be when they grow up? Lessons from epidermal patterning in Arabidopsis // Annu. Rev. Plant Biol. 2003. Т. 54. № 1. С. 403–430.
206. Larkin R. M. Tetrapyrrole signaling in plants // Front. Plant Sci. 2016. Т. 7. С. 1586.
207. Lee C.-C. и др. Mutation of a Nopp140 gene dao-5 alters rDNA transcription and increases germ cell apoptosis in C. elegans // Cell Death Dis. 2014. Т. 5. № 4. С. e1158–e1158.
208. Lee J. и др. Mutation of plastid ribosomal protein L13 results in an albino seedling-lethal phenotype in rice // Plant Breed. Biotechnol. 2019. Т. 7. № 4. С. 395–404.
209. Lefrançois L., Masopust D. The road not taken: Memory T cell fate «decisions» // Nat. Immunol. 2009. Т. 10. № 4. С. 369–370.
210. Lewis L. A. Hold the salt: Freshwater origin of primary plastids // Proc. Natl. Acad. Sci. U. S. A. 2017.

211. Li B., Dewey C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome // BMC Bioinformatics. 2011.
212. Li H., Ruan J., Durbin R. Maq: Mapping and assembly with qualities // Version 06. 2008a. Т. 3. С. 508.
213. Li H., Ruan J., Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores // Genome Res. 2008b.
214. Li N., Xu R., Li Y. Molecular Networks of Seed Size Control in Plants // Annu. Rev. Plant Biol. 2019. Т. 70. № 1. С. 435–463.
215. Li X. и др. A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of RNA-seq data // PLOS ONE. 2017.
216. Li Z. и др. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph // Brief. Funct. Genomics. 2012.
217. Liao Y., Smyth G. K., Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features // Bioinformatics. 2014.
218. Liebers M. и др. Regulatory shifts in plastid transcription play a key role in morphological conversions of plastids during plant development // Front. Plant Sci. 2017. Т. 8. С. 23.
219. Lin H. N., Hsu W. L. DART: A fast and accurate RNA-seq mapper with a partitioning strategy // Bioinformatics. 2018b.
220. Lindner R., Friedel C. C. A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq // PLoS ONE. 2012.
221. Liu S. C. и др. Transcriptomic analysis of tea plant responding to drought stress and recovery // PLoS ONE. 2016. Т. 11. № 1. С. 1–21.
222. Liu X. и др. WSL5, a pentatricopeptide repeat protein, is essential for chloroplast biogenesis in rice under cold stress // J. Exp. Bot. 2018. Т. 69. № 16. С. 3949–3961.
223. Long Z. и др. Genetic mapping and evolutionary analyses of the black grain trait in Barley // Front. Plant Sci. 2019.
224. Łopusiewicz Ł. Antioxidant, antibacterial properties and the light barrier assessment of raw and purified melanins isolated from Citrullus lanatus (watermelon) seeds // Herba Pol. 2018.
225. Love M., Anders S., Huber W. Differential analysis of RNA-Seq data at the gene level using the DESeq package // Eur. Mol. Biol. Lab. 2013.
226. Lu S. и др. CDD/SPARCLE: the conserved domain database in 2020 // Nucleic Acids Res. 2020. Т. 48. № D1. С. D265–D268.
227. Lu Y. и др. Promotion effects of flavonoids on browning induced by enzymatic oxidation of tyrosinase: structure–activity relationship // RSC Adv. 2021. Т. 11. № 23. С. 13769–13779.
228. Lu Y., Yao J. Chloroplasts at the Crossroad of Photosynthesis, Pathogen Infection and Plant Defense // Int. J. Mol. Sci. 2018. Т. 19. № 12. С. 3900.
229. Lundqvist U., Franckowiak J. D., Konishi T. New and revised descriptions of barley genes. , 1997.

230. Lyons S. M. и др. A subset of replication-dependent histone mRNAs are expressed as polyadenylated RNAs in terminally differentiated tissues // Nucleic Acids Res. 2016. Т. 44. № 19. С. 9190–9205.
231. Mantione K. J. и др. Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq // Med. Sci. Monit. Basic Res. 2014. Т. 20. С. 138–142.
232. Margulies M. и др. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors // Nat. Biotechnol. 2006. Т. 437. № 7057. С. 376–380.
233. Marioni J. C. и др. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays // Genome Res. 2008.
234. Martinez-Seidel F. и др. Systematic review of plant ribosome heterogeneity and specialization // Front. Plant Sci. 2020. Т. 11. С. 948.
235. Martínez-Zapater J. M. Genetic Analysis of Variegated Mutants in Arabidopsis // J. Hered. 1993. Т. 84. № 2. С. 138–140.
236. Mascher M. и др. Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley // Nat. Genet. 2016. Т. 48. № 9. С. 1089–1093.
237. Matus-Ortega M. G. и др. The alternative NADH dehydrogenase is present in mitochondria of some animal taxa // Comp. Biochem. Physiol. Part D Genomics Proteomics. 2011. Т. 6. № 3. С. 256–263.
238. Mavi K. The relationship between seed coat color and seed quality in watermelon Crimson sweet // Hortic. Sci. 2010.
239. Maza E. и др. Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: a matter of relative size of studied transcriptomes // Commun. Integr. Biol. 2013. Т. 6. № 6. С. e25849.
240. McCarron J. G. и др. Examining the role of mitochondria in Ca2+ signaling in native vascular smooth muscle // Microcirculation. 2013. Т. 20. № 4. С. 317–329.
241. McGettigan P. A. Transcriptomics in the RNA-seq era // Curr. Opin. Chem. Biol. 2013. Т. 17. № 1. С. 4–11.
242. Meyer E. H., Letts J. A., Maldonado M. Structural insights into the assembly and the function of the plant oxidative phosphorylation system // New Phytol. 2022. Т. 235. № 4. С. 1315–1329.
243. Mierziak J., Kostyn K., Kulma A. Flavonoids as important molecules of plant interactions with the environment // Molecules. 2014. Т. 19. № 10. С. 16240–16265.
244. Miflin B. J., Habash D. Z. The role of glutamine synthetase and glutamate dehydrogenase in nitrogen assimilation and possibilities for improvement in the nitrogen utilization of crops // J. Exp. Bot. 2002. Т. 53. № 370. С. 979–987.
245. Mistry J. и др. Pfam: The protein families database in 2021 // Nucleic Acids Res. 2020. Т. 49. № D1. С. D412–D419.
246. Mochizuki N. и др. Arabidopsis genomes uncoupled 5 (GUN5) mutant reveals the involvement of Mg-chelatase H subunit in plastid-to-nucleus signal transduction // Proc. Natl. Acad. Sci. 2001. Т. 98. № 4. С. 2053–2058.
247. Molina J. и др. Possible loss of the chloroplast genome in the parasitic flowering plant Rafflesia lagascae (Rafflesiaceae) // Mol. Biol. Evol. 2014. Т. 31. № 4. С. 793–803.

248. Molina-Cano J. L. и др. Morocco as a possible domestication center for barley: biochemical and agromorphological evidence // Theor. Appl. Genet. 1987.
249. Molina-Cano J. L. и др. Further evidence supporting Morocco as a centre of origin of barley // Theor. Appl. Genet. 1999. Т. 98. № 6. С. 913–918.
250. Molina-Cano J.-L. и др. Chloroplast DNA microsatellite analysis supports a polyphyletic origin for barley // Theor. Appl. Genet. 2005. Т. 110. № 4. С. 613–619.
251. Møller M. G. и др. Chlorophyll Biosynthetic Enzymes and Plastid Membrane Structures in Mutants of Barley (Hordeum vulgare L.) // Hereditas. 1997. Т. 127. № 3. С. 181–191.
252. Morrow I. C., Parton R. G. Flotillins and the PHB Domain Protein Family: Rafts, Worms and Anaesthetics: Flotillins and the PHB Family // Traffic. 2005. Т. 6. № 9. С. 725–740.
253. Necci M. и др. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins // Bioinformatics. 2017. Т. 33. № 9. С. 1402–1404.
254. Newton A. C. и др. Crops that feed the world 4. Barley: a resilient crop? Strengths and weaknesses in the context of food security // Food Secur. 2011. Т. 3. № 2. С. 141–178.
255. Nguyen C. V. и др. Tomato GOLDEN2-LIKE transcription factors reveal molecular gradients that function during fruit development and ripening // Plant Cell. 2014. Т. 26. № 2. С. 585–601.
256. Nikkanen L., Rintamäki E. Chloroplast thioredoxin systems dynamically regulate photosynthesis in plants // Biochem. J. 2019. Т. 476. № 7. С. 1159–1172.
257. Nonaka S. A new type of cultivar, Mitake, with very few in number, but thick and stiff culms // Barley Genet. Newsl. 1973. № 10. С. 47–51.
258. Nuell M. J. и др. Prohibitin, an Evolutionarily Conserved Intracellular Protein That Blocks DNA Synthesis in Normal Fibroblasts and HeLa Cells // MOL CELL BIOL. 1991.
259. O'Neil D., Glowatz H., Schlumpberge M. Ribosomal RNA depletion for efficient use of RNA-seq capacity // Curr. Protoc. Mol. Biol. 2013. № SUPPL.103. С. 1–8.
260. Osellame L. D., Blacker T. S., Duchen M. R. Cellular and molecular mechanisms of mitochondrial function // Best Pract. Res. Clin. Endocrinol. Metab. 2012. Т. 26. № 6. С. 711–723.
261. Pal A. K., Gajjar D. U., Vasavada A. R. DOPA and DHN pathway orchestrate melanin synthesis in Aspergillus species // Med. Mycol. 2014.
262. Pandey A. K., Dhakal M. R. Phytomelanin in compositae // Curr. Sci. 2001.
263. Patterson J. и др. Impact of sequencing depth and technology on de novo RNA-Seq assembly // BMC Genomics. 2019. Т. 20. № 1. С. 604.
264. Payá-Milans M. и др. Comprehensive evaluation of RNA-Seq analysis pipelines in diploid and polyploid species // GigaScience. 2018.
265. Paysan-Lafosse T. и др. InterPro in 2022 // Nucleic Acids Res. 2023. Т. 51. № D1. С. D418–D427.
266. Peer W. A., Murphy A. S. Flavonoids and auxin transport: modulators or regulators? // Trends Plant Sci. 2007. Т. 12. № 12. С. 556–563.
267. Peterhansel C. и др. Photorespiration // Arab. Book. 2010. Т. 8. С. e0130.

268. Peterhansel C., Maurino V. G. Photorespiration Redesigned // Plant Physiol. 2011. Т. 155. № 1. С. 49–55.
269. Pfannschmidt T. и др. Plastid RNA polymerases: orchestration of enzymes with different evolutionary origins controls chloroplast biogenesis during the plant life cycle // J. Exp. Bot. 2015. Т. 66. № 22. С. 6957–6973.
270. Pihlava J.-M. Identification of hordatines and other phenolamides in barley (Hordeum vulgare) and beer by UPLC-QTOF-MS // J. Cereal Sci. 2014. Т. 60. № 3. С. 645–652.
271. Pogson B. J., Ganguly D., Albrecht-Borth V. Insights into chloroplast biogenesis and development // Biochim. Biophys. Acta BBA-Bioenerg. 2015. Т. 1847. № 9. С. 1017–1024.
272. Ponce-Toledo R. I. и др. An early-branching freshwater cyanobacterium at the origin of plastids // Curr. Biol. 2017. Т. 27. № 3. С. 386–391.
273. Prina A. R. Mutator-induced cytoplasmic mutants in barley: genetic evidence of activation of a putative chloroplast transposon // J. Hered. 1996. Т. 87. № 5. С. 385–389.
274. Prina A. R. и др. A cytoplasmically inherited mutant controlling early chloroplast development in barley seedlings // Theor. Appl. Genet. 2003. Т. 107. № 8. С. 1410–1418.
275. Prota G. Recent advances in the chemistry of melanogenesis in mammals // J. Invest. Dermatol. 1980. Т. 75. № 1. С. 122–127.
276. Qin D. и др. Characterization and fine mapping of a novel barley Stage Green-Revertible Albino Gene (HvSGRA) by Bulked Segregant Analysis based on SSR assay and Specific Length Amplified Fragment Sequencing // BMC Genomics. 2015.
277. Qiu J. и др. Genome-wide analysis of the protein phosphatase 2C genes in tomato // Genes. 2022. Т. 13. № 4. С. 604.
278. Qiu Z. и др. The rice white green leaf 2 gene causes defects in chloroplast development and affects the plastid ribosomal protein S9 // Rice. 2018. Т. 11. № 1. С. 1–12.
279. Quaglieri A., Flensburg C., Speed T. P. Finding a suitable library size to call variants in RNA-seq.
280. Quinlan A. R., Hall I. M. BEDTools: A flexible suite of utilities for comparing genomic features // Bioinformatics. 2010.
281. Rajalingam K., Rudel T. Ras-Raf Signaling Needs Prohibitin // Cell Cycle. 2005. Т. 4. № 11. С. 1503–1505.
282. Rajkumar A. P. и др. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq // BMC Genomics. 2015.
283. Reyes D. и др. Overexpression of a protein phosphatase 2C from beech seeds in Arabidopsis shows phenotypes related to abscisic acid responses and gibberellin biosynthesis // Plant Physiol. 2006. Т. 141. № 4. С. 1414–1424.
284. Rice-Evans C. A., Miller N. J., Paganga G. Antioxidant properties of phenolic compounds // Trends Plant Sci. 1997.
285. Richardson L. G., Singhal R., Schnell D. J. The integration of chloroplast protein targeting with plant developmental and stress responses // BMC Biol. 2017. Т. 15. С. 1–10.

286. Ripoll J. J. и др. Antagonistic interactions between Arabidopsis K-homology domain genes uncover PEPPER as a positive regulator of the central floral repressor FLOWERING LOCUS C // Dev. Biol. 2009. Т. 333. № 2. С. 251–262.
287. Robertson D. W. и др. A Summary of Linkage Studies in Cultivated Barley, Hordeum Species: Supplement III, 1954–1963 1 // Crop Sci. 1965. Т. 5. № 1. С. 33–43.
288. Robertson G. и др. De novo assembly and analysis of RNA-seq data // Nat. Methods. 2010.
289. Robinson M. D., McCarthy D. J., Smyth G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. // Bioinforma. Oxf. Engl. 2010. Т. 26. № 1. С. 139–40.
290. Robles J. A. и др. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing // BMC Genomics. 2012. Т. 13. № 1.
291. Rodríguez V. M. и др. Genetic regulation of cold-induced albinism in the maize inbred line A661 // J. Exp. Bot. 2013.
292. Rodríguez-Cazorla E. и др. K-homology Nuclear Ribonucleoproteins Regulate Floral Organ Identity and Determinacy in Arabidopsis // PLOS Genet. 2015. Т. 11. № 2. С. 1–28.
293. Rodríguez-Concepción M., Boronat A. Elucidation of the methylerythritol phosphate pathway for isoprenoid biosynthesis in bacteria and plastids. A metabolic milestone achieved through genomics // Plant Physiol. 2002.
294. Rodríguez-Martín B. и др. ChimPipe: Accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data // BMC Genomics. 2017.
295. Romani I. и др. Versatile roles of Arabidopsis plastid ribosomal proteins in plant growth and development // Plant J. 2012. Т. 72. № 6. С. 922–934.
296. Rommens C. M. и др. Engineered native pathways for high kaempferol and caffeoylquinate production in potato // Plant Biotechnol. J. 2008. Т. 6. № 9. С. 870–886.
297. Rouet-Mayer M.-A., Ralambosoa J., Philippon J. Roles of o-quinones and their polymers in the enzymic browning of apples // Phytochemistry. 1990. Т. 29. № 2. С. 435–440.
298. Rumble S. M. и др. SHRiMP: Accurate mapping of short color-space reads // PLoS Comput. Biol. 2009.
299. Sadali N. M. и др. Differentiation of chromoplasts and other plastids in plants // Plant Cell Rep. 2019. Т. 38. С. 803–818.
300. Sáez-Vásquez J., Delseny M. Ribosome biogenesis in plants: from functional 45S ribosomal DNA organization to ribosome assembly factors // Plant Cell. 2019. Т. 31. № 9. С. 1945–1967.
301. Safi A. и др. The world according to GARP transcription factors // Curr. Opin. Plant Biol. 2017.
302. Sakamoto W. Leaf-variegated mutations and their responsible genes in Arabidopsis thaliana // Genes Genet. Syst. 2003. Т. 78. № 1. С. 1–9.
303. Schena M. и др. Quantitative monitoring of gene expression patterns with a complementary DNA microarray // Science. 1995.

304. Schertl P., Braun H.-P. Respiratory electron transfer pathways in plant mitochondria // Front. Plant Sci. 2014. Т. 5.
305. Schläpfer P. и др. Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants // Plant Physiol. 2017. Т. 173. № April. С. 2041–2059.
306. Schliesky S. и др. RNA-seq assembly - Are we there yet? // Front. Plant Sci. 2012.
307. Schmieder R., Edwards R. Quality control and preprocessing of metagenomic datasets // Bioinformatics. 2011. Т. 27. № 6. С. 863–864.
308. Schulz M. H. и др. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels // Bioinformatics. 2012.
309. Schwender J. и др. Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds // Nature. 2004. Т. 432. № 7018. С. 779–782.
310. See D. и др. Mapping Genes Controlling Variation in Barley Grain Protein Concentration // Crop Sci. 2002. Т. 42. № 3. С. 680–685.
311. Shendure J. The beginning of the end for microarrays? // Nat. Methods. 2008. Т. 5. № 7. С. 585–587.
312. Shi L.-X., Theg S. M. The chloroplast protein import system: from algae to trees // Biochim. Biophys. Acta BBA-Mol. Cell Res. 2013. Т. 1833. № 2. С. 314–331.
313. Shimizu T. и др. The retrograde signaling protein GUN1 regulates tetrapyrrole biosynthesis // Proc. Natl. Acad. Sci. 2019. Т. 116. № 49. С. 24900–24906.
314. Shmakov N. A. и др. Identification of nuclear genes controlling chlorophyll synthesis in barley by RNA-seq // BMC Plant Biol. 2016. Т. 16. № Suppl 3.
315. Shoeva O. Y. и др. Melanin formation in barley grain occurs within plastids of pericarp and husk cells // Sci. Rep. 2020.
316. Siedow J. N., Umbach A. L. The mitochondrial cyanide-resistant oxidase: structural conservation amid regulatory diversity // Biochim. Biophys. Acta BBA-Bioenerg. 2000. Т. 1459. № 2–3. С. 432–439.
317. Sims D. и др. Sequencing depth and coverage: Key considerations in genomic analyses // Nat. Rev. Genet. 2014. Т. 15. № 2. С. 121–132.
318. Siomi H. и др. The protein product of the fragile X gene, FMR1, has characteristics of an RNA-binding protein // Cell. 1993. Т. 74. № 2. С. 291–298.
319. Sitiwin E. и др. Shedding light on melanins within in situ human eye melanocytes using 2-photon microscopy profiling techniques // Sci. Rep. 2019.
320. Smid M. и др. Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons // BMC Bioinformatics. 2018.
321. Smith A. M. и др. Highly-multiplexed barcode sequencing: An efficient method for parallel analysis of pooled samples // Nucleic Acids Res. 2010. Т. 38. № 13. С. 1–7.
322. Smith P. M. C., Atkins C. A. Purine Biosynthesis. Big in Cell Division, Even Bigger in Nitrogen Assimilation // Plant Physiol. 2002. Т. 128. № 3. С. 793–802.
323. Solano F. Melanins: skin pigments and much more—types, structural models, biological functions, and formation routes // New J. Sci. 2014. Т. 2014. С. 1–28.
324. Soneson C., Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data // BMC Bioinformatics. 2013. Т. 14.

325. Sreenivasulu N., Graner A., Wobus U. Barley Genomics: An Overview // Int. J. Plant Genomics. 2008. Т. 2008. С. 486258.
326. Susek R. E., Ausubel F. M., Chory J. Signal transduction mutants of Arabidopsis uncouple nuclear CAB and RBCS gene expression from chloroplast development // Cell. 1993. Т. 74. № 5. С. 787–799.
327. Svensson J. T. и др. Transcriptome Analysis of Cold Acclimation in Barley Albina and Xantha Mutants // Plant Physiol. 2006. Т. 141. № 1. С. 257–270.
328. Takahashi R., Hayashi J. Linkage study of albino lemma character in barley // Ber Ohara Inst Landw Biol Okayama Univ. 1959. № 11. С. 132–140.
329. Taketa S. и др. Mutations in a Golden2-Like Gene Cause Reduced Seed Weight in Barley albino lemma 1 Mutants // Plant Cell Physiol. 2021. Т. 62. № 3. С. 447–457.
330. Tan C. и др. Characterization of genome-wide variations induced by gamma-ray radiation in barley using RNA-Seq // BMC Genomics. 2019. Т. 20. С. 1–8.
331. Tang H. и др. Kaempferol, the melanogenic component of Sanguisorba officinalis, enhances dendricity and melanosome maturation/transport in melanocytes // J. Pharmacol. Sci. 2021. Т. 147. № 4. С. 348–357.
332. Tang X. и др. A missense mutation of plastid RPS4 is associated with chlorophyll deficiency in Chinese cabbage (Brassica campestris ssp. pekinensis) // BMC Plant Biol. 2018.
333. Tarangini K., Mishra S. Production of melanin by soil microbial isolate on fruit waste extract: two step optimization of key parameters // Biotechnol. Rep. 2014. Т. 4. С. 139–146.
334. Thomas P. D. и др. PANTHER: Making genome-scale phylogenetics accessible to all // Protein Sci. 2022. Т. 31. № 1. С. 8–22.
335. Tian T. и др. AgriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update // Nucleic Acids Res. 2017. Т. 45. № W1.
336. Timm S. и др. The regulatory interplay between photorespiration and photosynthesis // J. Exp. Bot. 2016. Т. 67. № 10. С. 2923–2929.
337. Tombuloglu G. и др. High-throughput transcriptome analysis of barley (Hordeum vulgare) exposed to excessive boron // Gene. 2015. Т. 557. № 1. С. 71–81.
338. Toshoji H. и др. Effects of chloroplast dysfunction on mitochondria: white sectors in variegated leaves have higher mitochondrial DNA levels and lower dark respiration rates than green sectors // Protoplasma. 2012. Т. 249. С. 805–817.
339. Trapnell C., Pachter L., Salzberg S. L. TopHat: Discovering splice junctions with RNA-Seq // Bioinformatics. 2009.
340. Trick M. и др. Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing // Plant Biotechnol. J. 2009. Т. 7. № 4. С. 334–346.
341. Uberto R., Moomaw E. W. Protein Similarity Networks Reveal Relationships among Sequence, Structure, and Function within the Cupin Superfamily // PLOS ONE. 2013. Т. 8. № 9. С. e74477.
342. Varga M. и др. Structural characterization of allomelanin from black oat // Phytochemistry. 2016.

343. Varshney R. K. и др. A high density barley microsatellite consensus map with 775 SSR loci // Theor. Appl. Genet. 2007. Т. 114. № 6. С. 1091–1103.
344. Velculescu V. E. и др. Serial analysis of gene expression // Science. 1995.
345. Velculescu V. E. и др. Characterization of the yeast transcriptome // Cell. 1997.
346. Wang L. и др. DEGseq: An R package for identifying differentially expressed genes from RNA-seq data // Bioinformatics. 2009.
347. Wang L. и др. Observations on novel splice junctions from RNA sequencing data // Biochem. Biophys. Res. Commun. 2011a.
348. Wang L. и др. The arabidopsis purple acid phosphatase AtPAP10 is predominantly associated with the root surface and plays an important role in plant tolerance to phosphate limitation // Plant Physiol. 2011b.
349. Wang L. и др. Multifaceted roles of flavonoids mediating plant-microbe interactions // Microbiome. 2022. Т. 10. № 1. С. 233.
350. Wang L. F., Rhim J. W. Isolation and characterization of melanin from black garlic and sepia ink // LWT. 2019.
351. Wang S. и др. Prohibitin co-localizes with Rb in the nucleus and recruits N-CoR and HDAC1 for transcriptional repression // Oncogene. 2002. Т. 21. № 55. С. 8388–8396.
352. Wang S., Gribskov M. Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis // Bioinforma. Oxf. Engl. 2017.
353. Wang Y. и др. Origin of worldwide cultivated barley revealed by NAM-1 gene and grain protein content // Front. Plant Sci. 2015. Т. 6.
354. Wang Y. и др. Molecular evidence of RNA polymerase II gene reveals the origin of worldwide cultivated barley // Sci. Rep. 2016.
355. Wang Z., Gerstein M., Snyder M. RNA-Seq: A revolutionary tool for transcriptomics // Nat. Rev. Genet. 2009.
356. Waters B. M. и др. Mutations in Arabidopsis yellow stripe-like1 and yellow stripe-like3 reveal their roles in metal ion homeostasis and loading of metal ions in seeds // Plant Physiol. 2006. Т. 141. № 4. С. 1446–1458.
357. Weise S. E. и др. Transcriptional Regulation of the Glucose-6-Phosphate/Phosphate Translocator 2 Is Related to Carbon Exchange Across the Chloroplast Envelope // Front. Plant Sci. 2019. Т. 10.
358. Willcox G. The distribution, natural habitats and availability of wild cereals in relation to their domestication in the Near East: multiple events, multiple centres // Veg. Hist. Archaeobotany. 2005. Т. 14. № 4. С. 534–541.
359. Williams A. G. и др. RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis // Curr. Protoc. Hum. Genet. 2014.
360. Williams C. R. и др. Trimming of sequence reads alters RNA-Seq gene expression estimates // BMC Bioinformatics. 2016.
361. Williams C. R. и др. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq // BMC Bioinformatics. 2017.
362. Witte C.-P., Herde M. Nucleotide Metabolism in Plants1 [OPEN] // Plant Physiol. 2020. Т. 182. № 1. С. 63–78.

363. Woodson J. D., Chory J. Coordination of gene expression between organellar and nuclear genomes // Nat. Rev. Genet. 2008. Т. 9. № 5. С. 383–395.
364. Woodson J. D., Perez-Ruiz J. M., Chory J. Heme synthesis by plastid ferrochelatase I regulates nuclear gene expression in plants // Curr. Biol. 2011. Т. 21. № 10. С. 897–903.
365. Wu G.-Z. и др. Control of retrograde signaling by rapid turnover of GENOMES UNCOUPLED1 // Plant Physiol. 2018. Т. 176. № 3. С. 2472–2495.
366. Wu H., Wang C., Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data // Biostatistics. 2013.
367. Wu J. и др. Whole genome wide expression profiles of Vitis amurensis grape responding to downy mildew by using Solexa sequencing technology // BMC Plant Biol. 2010. Т. 10.
368. Wu T. D., Watanabe C. K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences // Bioinformatics. 2005.
369. Wu Z., Wang Z., Zhang K. Isolation and functional characterization of a glucose-6-phosphate/phosphate translocator (IbG6PPT1) from sweet potato (Ipomoea batatas (L.) Lam.) // BMC Plant Biol. 2021. Т. 21. № 1. С. 595.
370. Xiao Y. и др. Retrograde signaling by the plastidial metabolite MEcPP regulates expression of nuclear stress-response genes // Cell. 2012.
371. Xie Y. и др. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads // Bioinformatics. 2014.
372. Yagi Y., Shiina T. Recent advances in the study of chloroplast gene expression and its evolution // Front. Plant Sci. 2014. Т. 5.
373. Yan A. и др. The atypical histone variant H3.15 promotes callus formation in Arabidopsis thaliana // Development. 2020. Т. 147. № 11.
374. Yan Z. и др. Arabidopsis KHZ1 and KHZ2, two novel non-tandem CCCH zinc-finger and K-homolog domain proteins, have redundant roles in the regulation of flowering and senescence // Plant Mol. Biol. 2017. Т. 95. № 6. С. 549–565.
375. Yang C. и др. The impact of RNA-seq aligners on gene expression estimation. , 2015.
376. Yang Y. H. и др. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. // Nucleic Acids Res. 2002.
377. Yang Y., Smith S. A. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. // BMC Genomics. 2013.
378. Yao Z. Y., Qi J. H. Comparison of antioxidant activities of melanin fractions from chestnut shell // Molecules. 2016.
379. Yoshida K., Terashima I., Noguchi K. Distinct roles of the cytochrome pathway and alternative oxidase in leaf photosynthesis // Plant Cell Physiol. 2006. Т. 47. № 1. С. 22–31.
380. Yoshida K., Terashima I., Noguchi K. Up-regulation of mitochondrial alternative oxidase concomitant with chloroplast over-reduction by excess light // 2007.
381. Zhang T. и др. VIRESCENT-ALBINO LEAF 1 regulates leaf colour development and cell division in rice // J. Exp. Bot. 2018. Т. 69. № 20. С. 4791–4804.

382. Zhang Y. и др. Genome-Wide Characterization and Expression Analysis of KH Family Genes Response to ABA and SA in Arabidopsis thaliana // Int. J. Mol. Sci. 2022. Т. 23. № 1.
383. Zhao Z. и др. Identification of the Golden-2-like transcription factors gene family in *Gossypium hirsutum* // PeerJ. 2021. Т. 9. С. e12484.
384. Zhou K. и др. Albino seedling lethality 4; Chloroplast 30S ribosomal protein S1 is required for chloroplast ribosome biogenesis and early chloroplast development in rice // Rice. 2021. Т. 14. № 1. С. 1–12.
385. Zhou Q. и др. RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data // BMC Genomics. 2018. Т. 19. № 1. С. 144.
386. Zhu F. Anthocyanins in cereals: Composition and health effects // Food Res. Int. 2018.
387. Zhu H. и др. Expression patterns of purple acid phosphatase genes in Arabidopsis organs and functional analysis of AtPAP23 predominantly transcribed in flower // Plant Mol. Biol. 2005.
388. Zhu J. и др. Global dissection of alternative splicing uncovers transcriptional diversity in tissues and associates with the flavonoid pathway in tea plant (Camellia sinensis) // BMC Plant Biol. 2018. Т. 18. № 1. С. 1–12.
389. Zoschke R., Bock R. Chloroplast translation: structural and functional organization, operational control, and regulation // Plant Cell. 2018. Т. 30. № 4. С. 745–770.
390. Глаголева, А. Ю. Идентификация и анализ генов биосинтеза меланина в колосе ячменя (*Hordeum vulgare* L.) : дис. канд. биол. наук: 1.5.7 Генетика / Глаголева Анастасия Юрьевна – Н. 2022 – 147 с.

Дополнения

Дополнительная таблица 1. Гены, для которых была проведена верификация дифференциальной экспрессии с помощью количественной ПЦР в реальном времени. Приведены логарифмированные значения изменения уровней экспрессии, полученные с помощью от-ПЦР

Эксперимент	Ген	Лог. изменения экспрессии
Alm	HORVU3Hr1G039930	-3,28
	HORVU2Hr1G106880	-3,02
	HORVU2Hr1G068610	-3,36
	HORVU4Hr1G023580	-1,97
	HORVU2Hr1G040780	-4,20
	HORVU7Hr1G000900	-4,62
Blp	HORVU2Hr1G089440	2,34
	HORVU1Hr1G011930	4,77
	HORVU4Hr1G005920	2,34
	HORVU3Hr1G032370	2,36
	HORVU2Hr1G103000	4,70
	HORVU1Hr1G079200	-1,26
	HORVU4Hr1G087230	-5,44
	HORVU6Hr1G030390	2,02
	HORVU2Hr1G086380	0,06
	HORVU1Hr1G068030	0,68
	HORVU1Hr1G086680	0,70
	HORVU1Hr1G087070	-1,61
	HORVU2Hr1G103040	-0,51

Дополнительная таблица 2. Отнесение генов, локализованных в пластидном геноме ячменя, к трём функциональным группам (см. раздел 2.2.5 – функциональные анализ ДЭГ).

Ген	Описание	Группа
AGP50751	ATP synthase CF0 subunit I	прочие гены
AGP50750	ATP synthase CF0 subunit III	прочие гены
AGP50752	ATP synthase CF1 alpha subunit	прочие гены
AGP50762	ATP synthase CF1 beta subunit	прочие гены
AGP50761	ATP synthase CF1 epsilon subunit	прочие гены
AGP50779	ATP-dependent Clp protease proteolytic subunit	прочие гены
AGP50766	chloroplast envelope membrane protein	прочие гены
AGP50784	cytochrome b6	прочие гены

AGP50785	cytochrome b6/f complex subunit IV	прочие гены
AGP50767	cytochrome f	прочие гены
AGP50756	hypothetical chloroplast RF34	прочие гены
AGP50736	maturase K	прочие гены
AGP50798	NADH-plastoquinone oxidoreductase subunit 2	прочие гены
AGP50760	NADH-plastoquinone oxidoreductase subunit 3	прочие гены
AGP50801	NADH-plastoquinone oxidoreductase subunit 7	прочие гены
AGP50758	NADH-plastoquinone oxidoreductase subunit J	прочие гены
AGP50759	NADH-plastoquinone oxidoreductase subunit K	прочие гены
AGP50786	RNA polymerase alpha subunit	прочие гены
AGP50746	RNA polymerase beta	прочие гены
AGP50745	RNA polymerase beta subunit	прочие гены
AGP50747	RNA polymerase beta' subunit	прочие гены
AGP50789	translational initiation factor 1	прочие гены
AGP50765	photosystem I assembly protein Ycf4	гены фотосинтеза
AGP50755	photosystem I P700 apoprotein A1	гены фотосинтеза
AGP50780	photosystem II 47 kDa protein	гены фотосинтеза
AGP50741	photosystem II CP43 chlorophyll apoprotein	гены фотосинтеза
AGP50783	photosystem II phosphoprotein	гены фотосинтеза
AGP50735	photosystem II protein D1	гены фотосинтеза
AGP50740	photosystem II protein D2	гены фотосинтеза
AGP50739	photosystem II protein I	гены фотосинтеза
AGP50768	photosystem II protein J	гены фотосинтеза
AGP50738	photosystem II protein K	гены фотосинтеза
AGP50743	photosystem II protein M	гены фотосинтеза
AGP50781	photosystem II protein T	гены фотосинтеза
AGP50742	photosystem II protein Z	гены фотосинтеза
AGP50763	ribulose-1C5-bisphosphate carboxylase/oxygenase large subunit	гены фотосинтеза
AGP50791	ribosomal protein L14	гены белков рибосом
AGP50777	ribosomal protein L20	гены белков рибосом
AGP50794	ribosomal protein L22	гены белков рибосом
AGP50796	ribosomal protein L23	гены белков рибосом
AGP50775	ribosomal protein L33	гены белков рибосом
AGP50788	ribosomal protein L36	гены белков рибосом
AGP50787	ribosomal protein S11	гены белков рибосом

AGP50753	ribosomal protein S14	гены белков рибосом
AGP50800	ribosomal protein S15	гены белков рибосом
AGP50737	ribosomal protein S16	гены белков рибосом
AGP50776	ribosomal protein S18	гены белков рибосом
AGP50795	ribosomal protein S19	гены белков рибосом
AGP50748	ribosomal protein S2	гены белков рибосом
AGP50793	ribosomal protein S3	гены белков рибосом
AGP50757	ribosomal protein S4	гены белков рибосом
AGP50799	ribosomal protein S7	гены белков рибосом
AGP50790	ribosomal protein S8	гены белков рибосом

Дополнительная таблица 3. Приоретизация конвейеров биоинформатической обработки в эксперименте Alm.

Метод карт.	Метод фильт.	Метод поиска ДЭГ	Показатели				Ранги				Сумма рангов
			corr	stdev	%_ mapped	%_ uniq	corr	stdev	%_ mapped	%_ uniq	
dart	rnamap	edgeR	0,8382	0,0579	98,7740	87,4043	36	25	12	11	84
dart	rnamap	DEGseq	0,8363	0,0526	98,7740	87,4043	31	30	12	11	84
dart	unfil	DEGseq	0,8363	0,0526	98,7740	87,4043	32	31	10	10	83
dart	rnamap	DESeq2	0,8381	0,0582	98,8069	85,3196	35	23	12	11	81
dart	unfil	DESeq2	0,8381	0,0582	98,8069	85,3196	34	22	10	10	76
dart	chain	DEGseq	0,8164	0,0582	98,8069	85,3196	27	24	11	12	74
dart	unfil	edgeR	0,8375	0,0584	98,6988	79,0480	33	21	10	10	74
toph	unfil	DEGseq	0,8319	0,0448	98,6988	79,0480	30	36	3	1	70
star	chain	DEGseq	0,8070	0,0562	98,6988	79,0480	21	28	9	9	67
star	rnamap	DEGseq	0,8140	0,0563	64,5500	58,7600	25	26	8	8	67
star	unfil	DEGseq	0,8140	0,0563	64,5500	58,7600	26	27	7	7	67
toph	chain	DEGseq	0,8201	0,0503	64,5500	58,7600	28	34	2	3	67
toph	rnamap	DEGseq	0,8319	0,0448	59,7517	54,7000	29	35	1	2	67
hisat	unfil	DEGseq	0,8098	0,0525	59,7517	54,7000	24	33	5	4	66
hisat	rnamap	DEGseq	0,8098	0,0525	59,7517	54,7000	23	32	4	5	64
hisat	chain	DEGseq	0,8070	0,0529	59,7983	52,4867	22	29	6	6	63
star	rnamap	edgeR	0,8057	0,0691	59,7983	52,4867	20	14	8	8	50
star	unfil	edgeR	0,8055	0,0691	59,7983	52,4867	19	13	7	7	46

dart	chain	edgeR	0,7987	0,0718	78,8883	69,7250	14	8	11	12	45
star	rnamap	DESeq2	0,8049	0,0698	78,8883	69,7250	18	11	8	8	45
hisat	rnamap	edgeR	0,8001	0,0668	78,8883	69,7250	16	19	4	5	44
hisat	unfil	edgeR	0,8000	0,0663	74,4033	65,2433	15	20	5	4	44
dart	chain	DESeq2	0,7983	0,0723	74,4033	65,2433	11	7	11	12	41
star	unfil	DESeq2	0,8048	0,0698	74,4033	65,2433	17	10	7	7	41
hisat	rnamap	DESeq2	0,7986	0,0675	72,0650	62,0633	13	18	4	5	40
hisat	unfil	DESeq2	0,7986	0,0676	72,0650	62,0633	12	17	5	4	38
hisat	chain	edgeR	0,7920	0,0692	72,0650	62,0633	8	12	6	6	32
star	chain	edgeR	0,7896	0,0723	33,1905	29,2016	6	6	9	9	30
toph	rnamap	edgeR	0,7977	0,0677	33,1905	29,2016	10	16	1	2	29
hisat	chain	DESeq2	0,7896	0,0706	33,1905	29,2016	7	9	6	6	28
toph	unfil	edgeR	0,7973	0,0679	31,6817	27,8240	9	15	3	1	28
star	chain	DESeq2	0,7875	0,0735	31,6817	27,8240	3	5	9	9	26
toph	rnamap	DESeq2	0,7879	0,0739	31,6817	27,8240	5	4	1	2	12
toph	unfil	DESeq2	0,7878	0,0740	34,2236	25,6145	4	3	3	1	11
toph	chain	edgeR	0,7781	0,0775	34,2236	25,6145	2	2	2	3	9
toph	chain	DESeq2	0,7659	0,0848	34,2236	25,6145	1	1	2	3	7

Дополнительная таблица 4. Приоретизация конвейеров биоинформатической обработки в эксперименте Blp

		Метод поиска ДЭГ	Показатели				Ранги				Сумма рангов
Карт	Фильт.		corr	stdev	%_ mapped	%_ uniq	corr	std ev	%_ mapped	%_ uniq	
hisat	rnamap	edgeR	0,9297	0,0209	67,2333	59,9467	36	36	5	7	84
hisat	chain	edgeR	0,9295	0,0209	77,6417	68,3917	34	34	6	9	83
hisat	chain	DESeq2	0,9284	0,0211	77,6417	68,3917	31	31	6	9	77
hisat	unfil	edgeR	0,9296	0,0209	54,7750	34,4200	35	35	4	3	77
hisat	rnamap	DESeq2	0,9285	0,0211	67,2333	59,9467	32	32	5	7	76
star	chain	edgeR	0,9266	0,0217	90,2383	74,3900	28	28	9	10	75
hisat	unfil	DESeq2	0,9285	0,0211	54,7750	34,4200	33	33	4	3	73
star	rnamap	edgeR	0,9269	0,0217	84,2050	65,6133	29	29	7	8	73
star	unfil	edgeR	0,9271	0,0216	85,0933	34,6733	30	30	8	4	72
star	chain	DESeq2	0,9256	0,0220	90,2383	74,3900	25	25	9	10	69
star	rnamap	DESeq2	0,9258	0,0220	84,2050	65,6133	27	27	7	8	69
star	unfil	DESeq2	0,9258	0,0220	85,0933	34,6733	26	26	8	4	64
hisat	rnamap	DEGseq	0,9236	0,0229	67,2333	59,9467	24	21	5	7	57
dart	chain	edgeR	0,9200	0,0241	98,2820	82,1918	18	15	11	12	56
hisat	chain	DEGseq	0,9235	0,0229	77,6417	68,3917	22	19	6	9	56
star	chain	DEGseq	0,9217	0,0236	90,2383	74,3900	19	16	9	10	54
star	rnamap	DEGseq	0,9218	0,0235	84,2050	65,6133	20	17	7	8	52
star	unfil	DEGseq	0,9219	0,0235	85,0933	34,6733	21	18	8	4	51

dart	rnamap	edgeR	0,9197	0,0242	98,2733	74,5048	16	13	10	11	50
hisat	unfil	DEGseq	0,9236	0,0229	54,7750	34,4200	23	20	4	3	50
dart	unfil	edgeR	0,9198	0,0241	99,1242	40,5047	17	14	12	6	49
dart	chain	DESeq2	0,9190	0,0244	98,2820	82,1918	15	9	11	12	47
dart	rnamap	DESeq2	0,9188	0,0245	98,2733	74,5048	14	8	10	11	43
toph	chain	DESeq2	0,9180	0,0222	44,3667	36,8465	10	23	2	5	40
toph	rnamap	DESeq2	0,9180	0,0222	50,6000	15,8884	11	24	3	1	39
dart	chain	DEGseq	0,9148	0,0260	98,2820	82,1918	9	6	11	12	38
dart	unfil	DESeq2	0,9187	0,0245	99,1242	40,5047	13	7	12	6	38
toph	unfil	DESeq2	0,9180	0,0222	35,4500	24,3955	12	22	1	2	37
dart	rnamap	DEGseq	0,9147	0,0260	98,2733	74,5048	7	4	10	11	32
dart	unfil	DEGseq	0,9147	0,0260	99,1242	40,5047	8	5	12	6	31
toph	chain	DEGseq	0,9122	0,0243	44,3667	36,8465	4	10	2	5	21
toph	unfil	DEGseq	0,9122	0,0243	35,4500	24,3955	6	12	1	2	21
toph	rnamap	DEGseq	0,9122	0,0243	50,6000	15,8884	5	11	3	1	20
toph	chain	edgeR	0,9091	0,0264	44,3667	36,8465	3	3	2	5	13
toph	rnamap	edgeR	0,9091	0,0264	50,6000	15,8884	2	2	3	1	8
toph	unfil	edgeR	0,9087	0,0265	35,4500	24,3955	1	1	1	2	5