Отзыв официального оппонента

на диссертационную работу Шмакова Николая Александровича “Поиск генов, ассоциированных с частичным альбинизмом и меланизмом у ячменя *Hordeum vulgare* l., на основе анализа транскриптомных данных”, представленную на соискание ученой степени кандидата биологических наук по специальностям математическая биология, биоинформатика – 1.5.8 и генетика (биологические науки) – 1.5.7

Растительные пигменты представляют собой большую совокупность соединений, придающих окраску различным органам растений. Изучение пигментов является достаточно важной научной задачей. В связи с накоплением большого количества данных о геномах растений возникает задача идентификации генов, кодирующих ферменты, связанные с синтезом пигментов у немодельных организмов. Для этой цели может быть достаточно полезным изучение мутантных растений с изменённой цветовой окраской. В работе Николая Александровича был произведён поиск генов, ассоциированных с частичным альбинизмом и меланизмом, у ячменя *Hordeum vulgare* L.

У культурного ячменя белая окраска колосковой чешуи, цветковой чешуи (леммы), ушек и, частично, стебля и обусловлена мутацией в гене *Alm*. Про этот ген практически ничего неизвестно, хотя его мутантная модель была бы полезна для понимания механизмов работы ядерных и хлоропластных генов, участвующих в синтезе и распределении хлорофилла. Таким образом изучение мутантных растений с частичным альбинизмом (мутантов по гену *Alm*), так же должна помочь в понимании деталей механизмов координированного действия ядерных и хлоропластных генов.

Другим объектом исследования в данной работе стал мутант по гену *Blp*. Аллельный вариант этого гена обусловливает формирование черной окраски колосковой и цветковой чешуй и перикарпа ячменя. Основной вклад в окраску вызывается высоким содержанием алломеланинов – черных растительных пигментов полифенольной природы. Таким образом, изучение гена *Blp* может дать подходящую моделью для выявления ранее неизученных или слабо изученных метаболических и генных сетей растений, к каким относится путь синтеза алломеланинов.

Так же значительная часть работы и одним из ключевых результатов была разработка процедуры выбора конвейеров для обработки RNA-seq данных. Конвейер – это набор программных средств, запускаемых последовательно для выполнения определённой задачи. Для анализа RNA-seq данных разработано огромное количество программ для проведения оценки качества чтений, их картирования и т.д.. Таким образом имеется

огромное количество их различных комбинаций. В своей работе Николай Александрович предлагает процедуру выбора оптимального конвейера для анализируемых данных. Данная процедура является значимым биоинформатическим результатом и может быть интересна широкому кругу специалистов.

Диссертационная работа Шмакова Н.А. построена по классической схеме и состоит из списка использованных сокращений, введения, обзора литературы, раздела материалы и методы, результатов и их обсуждения, выводов, заключения, списка использованной литературы и дополнения. Диссертация изложена на 158 страницах и содержит 22 рисунка и 22 таблицы; список литературы содержит 390 источников. Материал диссертации изложен последовательно. Исследование проведено на высоком методологическом уровне, проделан большой объем работы. Надежность и достоверность полученных данных обеспечивается применением современных молекулярно-биологических и биоинформатических методов. Диссертационная работа носит полный и законченный характер – как в научном плане, так и в оформлении. Результаты диссертационной работы были опубликованы в четырёх зарубежных и российских научных журналах.

В обзоре литературы автор даёт информацию о ячмене и его экономических свойствах, так же приводит информацию о геноме ячменя. Далее автор даёт небольшое введение в биологию пластид и описывает меланизм и альбинизм растений. Большая часть обзора литературы посвящена обзору методов секвенирования второго поколения и описанию методов биоинформатического анализа использующихся при анализе RNA-seq данных. В разделе материалы и методы автор даёт описание используемого биологического материала. Так же приводится описание биоинформатических программных средств, использующихся при фильтрации и картировании библиотек, подсчёта уровня экспрессии, поиска дифференциальной экспрессии генов, функционального анализа ДЭГ, *de novo* реконструкции транскриптома, анализа *de novo* сборки транскриптома, а так же описывается метод для сравнения качества сборки транскриптомов.

Раздел результаты и их обсуждение состоит из двух больших смысловых частей

В первой части приводятся результаты анализа транскриптома почти изогенной линии ячменя i:Bw*Alm* в сравнении с Bowman. Анализ данных производился двумя путями. В первом случае использовался метод, основанный на картировании чтений. Как уже было сказано выше основной особенностью работы был подбор оптимального конвейера для обработки данных. В данной части работы выбор производился из 36 конвейеров обработки. В итоге оптимальным конвейером оказались два в которых для

картирования использовалась программа Dart, а для подсчёта ДЭГ использовались EdgeR или DEGseq. В итоге дифференциальная экспрессия была обнаружена только у семи генов, локализованных в районе *Alm*. Из них только один ген, кодирующий 40S рибосомный белок, повышает экспрессию в линии i:Bw*Alm* – его экспрессия повышена в 5 раз по сравнению с сортом Bowman. Гены с пониженной экспрессией в линии i:Bw*Alm* оказались ассоциированы с 11 метаболическими путями, преимущественно участвующими в фотосинтезе, фотодыхании, синтезе хлорофилла и усвоении азота, т.е. процессах, наиболее сильно нарушающихся при альбинизме растений. Для генов с повышенной экспрессией в линии i:Bw*Alm* значимо ассоциированных метаболических путей обнаружено не было.

Второй путь анализа представлял собой *de novo* сборку транскриптома. Было получено 5 сборок: Abyss, Spades, Trinity, Genome-guided trinity, а также метасборка. Итоговая выбранная сборка содержала 68414 контигов. Наибольшая длина контига в сборке была 9033 нуклеотида, средняя длина была 674 нуклеотида и N50 – 940 нуклеотидов. Была выбрана последовательность отсутствующая в транскриптоме сорта Bowman, но имеющая значимую экспрессию в i:Bw*Alm*. Белковый продукт ОРС этого транскрипта продемонстрировал сходство с последовательностью неаннотированного белка *H.vulgare* и прохибитин-1-подобным белком *Solanum pennellii*.

Во второй части приводятся результаты анализа транскриптома почти изогенной линии ячменя i:Bw*Blp* в сравнении с Bowman. Аналогично второй части анализ проводился с использованием картирования прочтений и составлением *de novo* сборки. При картировании чтений выбор конвейера производился так же из 36 вариантов. В итоге оптимальным конвейером оказался тот в котором для картирования использовалась программа Hisat2, а для подсчёта ДЭГ использовался EdgeR. В регионе *Blp* был обнаружен только один ген значимо изменяющий свою экспрессию. Этот ген кодирует фосфатазу пурпурной кислоты. Он повышает уровень экспрессии в линии i:Bw*Blp* в 39,8 раз по сравнению с сортом Bowman.

При *de novo* сборке транскриптома так же было получено несколько сборок и в результате из них была получена общая метасборка. Итоговая сборка содержала 32466 контигов. Было обнаружено 2 конига в линии i:Bw*Blp*, которые понижают экспрессию. Один из них оказался артефактом сборки, второй же имеет высокую гомологию к цитохром-P450-подобному белку. Ещё 2 контига имеют повышенную экспрессию в линии i:Bw*Blp*, для их пептидных продуктов наблюдается высокая гомология к фосфатазе белков 2C 68 и серин/треонин протеинкиназе PBL15 соответственно. И протеинкиназы, и

протеинфосфатазы выполняют самые разнообразные роли в растительных клетках, в том числе регулируют ответ на различные виды стресса, а также участвуют в формировании реакции на некоторые фитогормоны.

Дополнительно хочется отметить, что в данной работе показано, что для разных наборов данных могут потребоваться разные конвейеры обработки. В случае анализа данных линии i:Bw*Alm* оптимальным картировщиком оказался Dart, в случае же i:Bw*Brt* HiSat2. Картирование чтений один из базовых этапов анализа и ожидается, что на его выбор скорее должны влиять технические факторы, чем характер данных. Но данная работа показывает, что это не так и даже для данных, полученных с использованием одной и той же технологии секвенирования могут потребоваться разные наборы программ для анализа.

Диссертационная работа не лишена ряда недостатков:

1) В тексте диссертации были обнаружены следующие опечатки:

а) на стр 13 шестая строка «культру» нужно заменить на «культуру»

б) на стр. 29 во второй строке снизу пропущен предлог “с” между “20 миллиардов парных прочтений длинами до 250” “должно быть 20 миллиардов парных прочтений с длинами до 250”

в) на стр. 63 четвертая строка снизу «выровнены» на «выравнены»

2) В литературном обзоре упоминается понятие «функциональное питание». Что в данном контексте под ним подразумевается?

3) В литературном обзоре при описании графов де Брёйна не указано, как определяются их рёбра.

4) Насколько сильно геномная линия ячменя Bowman отличается от Morex. Могло ли это повлиять на качество анализа?

5) Проводилось ли для субоптимальных подходов к анализу данных поиск обогащённых GO терминов и метаболических путей. Насколько результаты были отличны от оптимального подхода к обработке?

6) Осуществлялась ли оценка скорости для 36 вариантов обработки? И насколько быстры были оптимальные подходы?

Эти замечания, однако, не снижают хорошего впечатления от работы и носят характер пожеланий или технических замечаний.

Диссертационная работа Шмакова Николая Александровича «Поиск генов, ассоциированных с частичным альбинизмом и меланизмом у ячменя Hordeum vulgare l.,

на основе анализа транскриптомных данных» соответствует требованиям, предъявляемым ВАК РФ к кандидатским диссертациям. Диссертация соответствует требованиям пп. 9,10,11,13,14 “Положения о присуждении ученых степеней” (утверждено Постановлением Правительства РФ от 24.09.2013 №842) для ученой степени кандидата наук, а ее автор заслуживает присуждения искомой ученой степени кандидата биологических наук по специальностям 1.5.8 – “Математическая биология, биоинформатика” и 1.5.7 – “Генетика (биологические науки)”.

Официальный оппонент,

Старший научный сотрудник лаборатории №19

Института проблем передачи информации им.

А.А. Харкевича,

кандидат физико-математических наук А.С. Касьянов

16 января 2024 года

Подпись
Касьянов А.С.
УДОСТОВЕРЯЮ
Зав. канцелярией