ОТЗЫВ

официального оппонента на диссертацию Комышева Евгения Геннадьевича на тему: «Разработка методов автоматического определения количественных характеристик, описывающих фенотипические признаки колоса пшеницы», представленной на соискание ученой степени кандидата биологических наук по специальности 03.01.09 – математическая биология, биоинформатика.

Актуальность избранной темы.

Увеличение производства продовольствия для растущего населения Земли является одной из глобальных проблем. При этом исследования изменения климата показали, что повышение урожайности будет затруднено по причине расширения засушливых и жарких районов. Таким образом, возрастает потребность в новых неприхотливых сортах с высокой продуктивностью и качеством урожая. Одной из важнейших сельскохозяйственных культур является пшеница, продуктивность которой связана с размером и формой колоса. В этой связи актуальной задачей становится создание высокопроизводительных методов фенотипирования на основе компьютерной обработки изображений и эффективных систем хранения результатов с возможностью удобного доступа к ним.

Данная работа служит ярким примером грамотного подхода к разработке и использованию таких методов и систем, сочетающих мобильное приложение и информационную систему с веб-интерфейсом. Использование развитых алгоритмов обработки изображений, определения фенотипических параметров, а также современных веб-технологий и баз данных открывает новые возможности для исследователей и селекционеров.

Научная новизна исследования, полученных результатов, выводов и рекомендаций, сформулированных в диссертации. К наиболее важным научным результатам исследования, характеризующим его новизну, могут быть отнесены:

- Впервые создано мобильное приложение SeedCounter для устройств под управлением ОС Android для подсчета зерен и определения их размеров.
- Впервые предложен новый метод автоматического определения количественных морфометрических характеристик колоса пшеницы на основе анализа цифровых двухмерных изображений, который позволяет описывать форму колоса на основе модели двух четырехугольников.
- Проведена оценка сходства и различий формы колосьев с использованием количественных характеристик, полученных у 14 образцов растений пяти генотипов мягкой пшеницы на основе параметров модели, оцененных путем анализа двухмерных цифровых изображений.
- Разработана компьютерная система SpikeDroid, реализующая технологии хранения разнородной информации, для поддержки работы селекционера по сбору данных по морфометрическим характеристикам колосьев пшеницы и их диких сородичей.

Значимость для науки и практики полученных автором результатов. Теоретическая значимость исследования состоит в том, что:

- Предложенные методы позволяют оценивать такие характеристики колоса пшеницы, как длина, ширина, остистость, количество зерен в колосе, а также характеристики зерен пшеницы (длина, ширина, проецируемая на поверхность площадь, коэффициенты округлости, закругленности, шероховатости и твердости формы контура и др.). Это дает возможность оценивать на основе анализа изображений характеристики урожайности растений с высокой степенью точности. Предложенные методы позволяют исключить субъективизм присущий человеку при проведении измерений, т.к. обработка цифровых изображений производится автоматически. Кроме того, предложенный подход не требуют от пользователя специфических знаний в области фенотипирования пшеницы.
- Собранные воедино в системе SpikeDroid данные с анализируемых изображений, ручного фенотипирования и генотипов позволяют повысить эффективность существующих методов сравнительной генетики пшениц, таких как ги-

бридологический метод анализа, метод возвратных скрещиваний, метод циклических скрещиваний и др.

Практическая значимость исследования заключается в том, что мобильное приложение SeedCounter позволяет выполнять измерения в полевых условиях, сохранять на мобильном устройстве данные, а также отправлять их на сервер посредством сети Интернет, когда он доступен. Разработанная система SpikeDroid позволяет существенно ускорить процесс массового фенотипирования благодаря автоматизации этапов начиная от получения изображений, заканчивая статистическим анализом занесенных в базу данных параметров.

Преимущества разработанных методов позволяют существенно повысить эффективность селекционно-генетических экспериментов в направлении создания новых высокопродуктивных сортов и линий пшеницы.

Степень обоснованности научных положений, выводов и рекомендаций, сформулированных в диссертации. Работа Комышева Е.Г. построена по общепринятой схеме и состоит из введения, обзора литературы, четырех глав, описывающих материалы и методы, результаты собственных исследований и их обсуждение, а также заключения, выводов и списка литературы. Список литературы включает в себя 135 источников, из них только 17 на русском языке, остальные 118 на английском языке, что говорит о хорошем владении автором зарубежной литературой. Из общего числа 10 источников или примерно 8,5% относятся к последним 5 годам (2017-2021 гг.). Работа изложена на 146 страницах и содержит 17 таблиц и 36 рисунков.

Структура и логика изложения соответствуют поставленным в диссертации задачам исследования, нацеленным на разработку методов автоматического определения количественных морфометрических характеристик колосьев и зерен пшеницы на основе анализа их цифровых изображений:

1. Разработка метода морфометрии зерен пшеницы с использованием мобильных устройств.

2. Разработка методов автоматического определения количественных характеристик формы и размера колоса на основе двухмерных изображений и его апробация на примере анализа колосьев пяти видов гексаплоидных пшениц.
3. Разработка базы данных для накопления, хранения и систематизации информации о фенотипических признаках колоса пшеницы.

Для решения поставленных задач автор опирается на обширную теоретическую и методологическую базу. Во введении обоснована актуальность темы исследования, сформулированы цель и задачи исследования, охарактеризованы научная новизна, теоретическая и практическая ценность исследования, сформулированы три положения, выносимые на защиту, содержащие наиболее существенные результаты работы, обладающие научной новизной.

Обзор литературы (стр. 16) очень подробен и обстоятелен, охватывает как описание объекта исследования и генома пшеницы (стр. 16-19), структуру растения и колоса пшеницы (стр. 19-25), гены, определяющие морфологию колоса (стр. 25-29), признаки пшеницы, связанные с урожайностью (стр. 29-32), методы обработки биологических цифровых изображений (стр. 33-36), сегментацию, морфологические преобразования и поиск контуров (стр. 36-44), так и описание компьютерных средств — библиотек анализа цифровых изображений, таких как Матлаб (стр. 45-46), среда R (стр. 46-47), библиотека OpenCV (стр. 47-51), ImageJ (стр. 47), scikit-image (стр. 48), а также описание методов фенотипирования колосьев (стр. 48) и анализа формы зерен (стр. 49-50). Уделено внимание использованию мобильных устройств (стр. 50-51), базам данных (стр. 51-52), феномике (стр. 52-53), онтологиям, таким как Gene Ontology, Plant Ontology и Crop Ontology (стр. 53-57), а также программам для статистического анализа данных (стр. 57-59). Приводиться собственное заключение (стр. 59-60).

В Главе 2 (стр. 61) описаны растительный материал и использованные в работе методы. Описаны генотипы пшеницы использованные для морфометрии зерен (стр. 61) и морфометрии колосьев (стр. 61-63), описаны условия выращивания. Приведены краткие характеристики методов обработки изображений (стр. 63-

65). Рассмотрена общая блок-схема предложенных алгоритмов обработки (стр. 65). В методах оценки точности алгоритмов анализа изображений описан индекс Жаккара (стр. 66), оценки ошибок MAE и MAPE (стр. 67-68), в методах статистического анализа — методы из пакетов R и PAST (стр. 68). Дана краткая характеристика разработки приложения для Android (стр. 69) и базы данных с веб-интерфейсом (стр. 69-70).

В Главе 3 (стр. 71) подробно рассмотрен метод и мобильное приложение для морфометрии зерен пшеницы с помощью мобильных устройств. Отмечается важность освещения, даны рекомендации для съемки в помещении и вне, а также для использования листов бумаги (стр. 71-72). Алгоритм обработки (стр. 72-75) реализован с помощью библиотеки OpenCV и включает распознавание листа бумаги (стр. 73-74), распознавание контуров зерен и аппроксимацию эллипсоидами (стр. 74-75). Дано описание интерфейса мобильного приложения SeedCounter (стр. 75-77). Точность работы алгоритма (стр. 77-87) была оценена по нескольким параметрам: количество идентифицированных зерен при разном освещении, длина и ширина отдельных зерен, сравнение с приложением SmartScan — ближайшим аналогом. Тесты показали, что средняя относительная ошибка подсчета числа зерен не высока, и составляет 2% относительно общего числа зерен, а абсолютная ошибка в среднем равна 1 зерну. Коэффициенты корреляции между реальным размером длины зерен (измеренное вручную) и её программной оценкой во всех экспериментах были не ниже 0,79 (коэффициент корреляции Пирсона, $p<0,01$). В заключении главы 3 сделан вывод о возможности использования метода в массовых селекционно-генетических экспериментах (стр. 87).

Главе 4 (стр. 88) посвящена применению метода морфометрии колоса пшеницы. Описаны два протокола получения цветных двумерных изображений: «на прищепке» и «на столе» (стр. 88-89). Приведены шаги идентификации колоса и остей на изображении: предварительная фильтрация с размытием по Гауссу (стр. 89), распознавание цветовой шкалы (стр. 89-90), сегментация (стр. 90-91),

идентификация остей (стр. 91-92) с подбором параметров методом оптимизации (стр. 93-94), идентификация контура колоса (стр. 94), построение осевой линии колоса итерационным алгоритмом (стр. 94-96), выпрямление колоса (стр. 96), подсчет интегральных характеристик формы — длины, площади, отношение площади к квадрату длины, округлость, индекс закругленности, индекс шероховатости и компактности (стр. 96-97). Представлена модель четырехугольников для контура колоса (стр. 97-99). Была проведена оценка точности распознавания областей остей и колоса (стр. 100-106), в которой среднее значение индекса Жаккара для тела колоса составило 0,925 и 0,932 на тестовой и обучающей выборках, соответственно, а для остей — 0,660 и 0,679, соответственно. Анализ параметров остистости (стр. 106-107) показал, что самую большую дисперсию площади остей имеют колосья с типом остистости “коротко-остые”, которые имеют как малые, так и большие площади остей. Анализ характеристик формы колоса (стр. 107-113) для четырнадцати генотипов пяти видов гексаплоидных пшениц (всего 160 образцов) показал, что наиболее вариабельными признаками являются характеристики площадей и периметр контура колоса. Анализ корреляций между характеристиками с помощью иерархической кластеризации (стр. 113-114) выявил, что в предложенном подходе форма колоса характеризуется тремя основными сегментами: основание, центральная часть и вершина. Анализ вариабельности морфометрических характеристик колоса (стр. 114-120) с помощью метода главных компонент показал, что виды *T. aestivum* и *T. compactum*, вид *T. spelta* и виды *T. antiquorum* и *T. sphaerococcum* разделяются по получаемым параметрам. В заключении (стр. 120-122) сделан вывод о том, что полученная в результате анализа информация о морфометрических характеристиках колоса может быть применена в высокопроизводительном, автоматизированном фенотипировании и при проведении селекционных экспериментов.

В Главе 5 (стр. 123) описана информационная система SpikeDriod, начиная с модели данных (стр. 123-125), включающей 5 таблиц и 4 отношения, далее к

использованным интернет-технологиям (стр. 125-126), пользовательскому интерфейсу (стр. 126-129), использующему QR-код, вмещающий до 4296 символов, заканчивая информационным содержанием (стр. 129-130) из 380 растений и 1475 фотографий колосьев. В заключение главы (стр. 130-131) указан адрес доступа к системе.

В заключении диссертационной работы (стр. 132) кратко обобщены основные результаты исследования. По результатам работы сформулированы обоснованные выводы (стр. 134).

Таким образом, на основе достаточного анализа предметной области, адекватной постановки научной проблемы и задач исследования, корректного применения современных подходов и методов получены вполне достоверные и обоснованные результаты.

Замечания по содержанию и оформлению диссертации.

В словаре сокращений на стр. 6 имеется неточность в употреблении терминов, а именно приводятся одинаковые расшифровки для SNP и SNV, однако следовало уточнить, что вариант с одним нуклеотидом (SNV) может встретиться у одного индивида, но станет однонуклеотидным полиморфизмом, если будет встречаться в популяции с некоторой частотой.

Имеются неточности в приведенных ссылках. Например, на стр. 27 для Huang et al не указан год (2020), а на стр. 46 и 58 в ссылке на R — слово «Team» выглядит как фамилия, хотя это просто слово «команда» из сочетания «R Team», и «et al» не нужно. Ссылку на ПО МАТЛАБ (стр. 45) следовало указать на сайт компании Mathworks.

Встречаются не очень удачные и, поэтому, не совсем понятные фразы. Например, на стр 27 автор пишет: «ALI-1 транскрипционно подавляет гены, расположенные ниже по течению, снижает содержание цитокининов и одновременно сдерживает передачу сигнала, что приводит к уменьшению количества клеток в ости.», однако о каком сигнале и течении идет речь не ясно. На стр. 37 упомянуты «непрерывные точки», на стр. 39 сказано: "эрозия удаляет белые шумы, но также удаляют точки объект переднего плана", что не очень понятно. На стр. 72 автор поясняет, что «устройство получения изображений (фотокамера или мобильное устройство) располагается на расстоянии около 50 см перпендикулярно над листом.», однако что чему перпендикулярно не ясно, логично было бы что телефон параллелен листу.

Имеются недоработки в формулах. Например, на стр. 29, формула 1, использована "*", скорее всего для обозначения умножения, однако такое обозначение арифметической операции больше в работе не встречается, корректнее умножение обозначать точкой. В формуле 2 на стр. 35 не подписаны обозначения, на стр. 99 в формулах для AI цифра 2 скорее всего степень, на это стоило пояснить.

В обзоре методов обработки изображений отсутствуют некоторые детали. Например, среди вариантов связности на стр. 36 не упомянут третий вариант: два пикселя либо соседи по связности крест, либо соседи по диагонали, если по связности крест нет пикселя с нужным уровнем, т.е. пиксели будут соседями по диагонали только если нет прямых соседей (см. Гонсалес и Вудс, «Цифровая обработка изображений»). При описании морфологических операций (стр. 38) сказано, что они обычно выполняются на двоичных изображениях, однако такие операции также интересны на полутоновых изображениях. Не упомянуты такие операции, как реконструкция, утолщение, поиск доменов и локальных максимумов.

При описании генетического алгоритма (стр. 41) следовало упомянуть операции мутации и селекции. Также стоило упомянуть такие продвинутые алгоритмы оптимизации как эволюционные стратегии и разностную эволюцию.

стр 43 стоило пояснить что такое простой классификатор. Нейронная сеть не является сложным объектом, сложности возникают при обучении. Нелинейная регрессионная модель или система ДУЧП сложнее, но в ней может быть меньше параметров и лучше описаны процессы, поэтому обучить ее - найти параметры - легче. Производительность ИНС достигается с помощью ГПУ, в том числе в с/х.

В описании программного обеспечения присутствуют неточности. Среди функций МАТЛАБ можно было указать и разные биоинформатические пакеты, и Simulink как средство визуального моделирования, и символьные вычисления, хотя перечисление всех функций Матлаба в данной работе выглядит лишним, т.к. к предмету относятся только методы обработки изображений, а повторение функций Матлаба на стр. 57 излишне. Сравнение библиотеки opencv (стр. 47) в одном ряду с Матлабом, R, ImageJ не совсем корректно. Кроме того, в списке не хватает представителей коммерческих программ — Imaris, например. Для ПО PAST (стр. 59) требуется ссылка на сайт или другой источник.

В описании разработки ПО для Android есть противоречивая информация: на стр. 69 указана версия API 15, стр. 87 — 14. Следовало также указать версии MySQL, Drupal, PHP, CentOS и конфигурацию сервера, использованных для разработки базы данных и веб-интерфейса. Также надо упомянуть противоречие в том, что интерфейс приложения и веб-интерфейс используют английский язык, но запятую для десятичной дроби, и отзывы также все оставлены на русском языке. Кроме того, Android Play Store (стр. 87) 6 марта 2012 года переиме-

нован в Google Play, где среди авторов приложения указан только Михаил Генаев.

Приведенные результаты тестирования мобильного приложения по точности (стр. 78) и времени (стр. 84) были получены на устаревших устройствах, первое из которых работает под управлением не поддерживаемой версии Android 2.3 (таб. 4). Кроме того, следовало указать по какой формуле cvtColor переводит RGB в оттенки серого, по какой формуле работает adaptiveThreshold, как подбираются параметры Кэнни (стр. 73).

На стр. 100 автор указывает, что использование цветокоррекции не оказало существенного влияния на точность, однако для проверки значимости учета цветокоррекции следовало бы привести стандартные отклонения индексов. Также непонятно сколько проводилось оптимизаций, на скольких разбиениях на обучающую и тестовую выборки. Если была проведена кросс-валидация, например со 100 случайными разбиениями, то хотелось бы видеть гистограммы результатов для обучающей и тестовой выборок. Тогда можно было бы сравнить средние или медианы по критерию, сделать вывод о различиях, выбрать лучшую модель (или ансамбль из нескольких моделей), проверить его точность на проверочной выборке, которую не использовали ни для обучения ни для тестирования. Также интересно было бы посмотреть на оптимальные значения 7 параметров, полученные генетическим алгоритмом. На рис. 25, стр. 103 по оси ординат отложен не процент либо все величины меньше 1%. Приведенные результаты показывают, что стоит брать разные параметры для разных масштабов.

На стр. 106 на основе рис. 26 автор делает вывод о том, что масштаб изображения значимо влияет на точность распознавания как тела колоса, так и остей, тип протокола оказывают существенное влияние на точность определения тела ко-

лоса, но не остей. Однако, значимость влияния надо оценивать по критерию сравнения медиан для разных условий, например Манна-Уитни, на глаз вывод не очевиден совсем.

На стр. 106 указано, что «распределение представлено на рис.27», однако рисунок подписан гистограмма, что больше соответствует действительности. Стоило нарисовать огибающие кривые для демонстрации распределений, либо сделать фиттинг каких-то распределений.

На стр. 113 мера близости генотипов может стать больше 1 или все корреляции должны быть больше 0, в любом случае стоило их привести. На рис.28 есть отрицательное сходство, стоило пояснить что это означает.

На стр. 115, рис. 29 зеленые линии для характеристик плохо различимы, поэтому вклад характеристик в главные компоненты не понятен, также не указано как компоненты объясняют дисперсию. Хотелось бы понимать какие признаки попали в какую компоненту. На стр. 116, рис. 31 повтор рис. 30, т.е. рисунок отношения длины колоса и площади в модели четырехугольников отсутствует.

По измеренным характеристикам логично было построить классификатор, например, на основе случайного леса из упоминавшегося в обзоре пакета scikit, что позволило бы также оценить значимость каждой характеристики при классификации.

По Глава 5 (стр. 124) остается непонятно, почему в базе хранятся не все характеристики колоса, которые ранее изучались, и не изучались другие характеристики, которые там хранятся, например, опушения. Также, стоило пояснить для создания системы осуществлялось написание кода или только настройка в административном интерфейсе Drupal (стр. 125). Кроме того, остается вопрос что

будет при превышении числа растений 4296, что не так и много для массового использования. По рис. 36 (стр. 130) стоило привести какой-то вывод, разделения на группы незаметно.

Также замечены опечатки:

Стр 9 связаны с ... количеств~~а~~**ом**

Стр 35 "~~Методы фильтрации шума на изображении разнятся в зависимости от типа искажения. Наиболее часто встречающиеся искажения на изображении:~~

Методы фильтрации шума на изображении разнятся в зависимости от типа искажения. Наиболее часто встречающи**м**ся приемом улучшения качества изображения является его размытие."

Стр 41 служит генетически~~х~~**й** алгоритм

Стр 46 R в основном используется для ... работ~~ой~~**ы**

Стр 44 Метод**,** позволяющий производить оценку биомассы

Стр 48 В особую категорию можно отнести так называемые неразрушающие методы изме~~н~~**р**ения

Стр 66 на основе двух тип~~ах~~**ов** оценок

Стр 83 Это может ~~быть~~ объясняться

Стр 108 Так параметры**,** определяющие -- запятая

Отмеченные недостатки, впрочем, не являются критическими, не умаляют ценности работы и не влияют на общую положительную оценку диссертационной работы Комышева Е.Г.

Заключение о соответствии диссертации критериям, установленным Положением о порядке присуждения ученых степеней. Диссертационное исследование Комышева Е.Г. выполнено на актуальную тему, представляет собой законченную научную работу, имеет теоретическую и практическую ценность.

Основные результаты диссертации изложены в 17 научных работах, из них пять — в рецензируемых научных журналах, входящих в перечень ВАК по специ-

альности 03.01.09 и десять работ в сборниках тезисов конференций, апробированы научной общественностью и получили положительную оценку. Получено два авторских свидетельства.

Автореферат и публикации соискателя отражают основное содержание диссертации.

Таким образом, диссертация Комышева Евгения Геннадьевича является научно-квалификационной работой, характеризуется новизной, содержит решение актуальной задачи, имеющей теоретическое и практическое значение, что соответствует требованиям п. 9 «Положения о порядке присуждения ученых степеней», утвержденного постановлением Правительства Российской Федерации от 24.09.2013 г. № 842, предъявляемым к диссертациям на соискание ученой степени кандидата биологических наук по специальности 03.01.09 – математическая биология, биоинформатика, а ее автор заслуживает присуждения искомой ученой степени.

Доцент ВШПМиВФ ИПММ,
федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский политехнический университет Петра Великого»,
кандидат биологических наук

К.Н. Козлов

Козлов Константин Николаевич
195251, Санкт-Петербург, Политехническая ул., д. 29, кор. НИК (АФ), пом. А.3.28,
Телефон: +7-812-290-9642, +7-904-518-8581
Адрес электронной почты: kozlov_kn@spbstu.ru
Место и адрес работы: ВШПМиВФ ИПММ ФГАОУ ВО «СПбПУ»
Должность: доцент, к.б.н

УДОСТОВЕРЯЮ
Ведущий специалист
по кадрам
«17» 05 2021 г.