WheatPGE – компьютерная система для анализа взаимосвязи признаков фенотипа, генотипа и параметров окружающей среды у пшеницы*

© М.А. Генаев[1], А.В. Дорошков[1], Д.А. Афонников[1,2]

[1]Институт цитологии и генетики СО РАН, Новосибирск

[2]Новосибирский государственный университет

mag@bionet.nsc.ru

Аннотация

Для решения задачи интеграции генотипических и фенотипических данных, а также параметров окружающей среды и анализа взаимосвязей между генотипом и фенотипом у пшеницы представлена система WheatPGE. Система служит для интеграции разнородных данных о растении, хранении и доступе к информации об отношениях, описывающих различные характеристики растения, его генотипа, фенотипа и факторов внешней среды. Система имеет простой и удобный веб-интерфейс и доступна по адресу www.wheatdb.org [1].

1 Введение

Современная биология характеризуется взрывным ростом данных в самых разных областях этой науки. Методы секвенирования ДНК и выявления полиморфизма генома позволяют быстро и эффективно устанавливать генотип, мутантные аллели для большого числа генов для тысяч организмов [2]. Эти достижения позволяют революционизировать методы селекции растений с новыми важными для сельского хозяйства признаками, что особенно актуально для таких широко используемых в сельском хозяйстве растений, как пшеница. Наряду с этим разрабатываются новые методы высокопроизводительного фенотипирования растений, позволяющие получать эквивалентные по объему массивы данных о фенотипических признаках [3 – 5]. Сопоставление большого количества таких данных позволит биологам получать новые знания о взаимосвязи между генотипом и фенотипом организмов [6]. Однако при решении этой задачи возникает проблема интеграции большого объема данных о фенотипах и генотипах растений, а также об условиях окружающей среды, с целью их дальнейшего анализа.

Для решения этой проблемы создаются различные базы данных. Например, проект GrainGenes [7] содержит информацию о генах аллелях и генетических маркерах различных злаков. Chlroloplast 2010 [8] аккумулирует информацию о различных морфологических признаках арабидопсиса. Также создаются инструменты, помогающие биологу-селекционеру. Эти базы данных, однако, не позволяют описать параметры генотипа, фенотипа для одного растения. Этот недостаток делает невозможным сбор материала для отдельных растений с целью его дальнейшей статистической обработки.

PlantDB [9] – инструмент на основе Microsoft Access для занесения базовой информации о генотипе и некоторых фенотипических признаках исследуемых растений. Эта база данных, в отличие от предыдущих, ориентирована на описание параметров каждого растения, для которого проводится эксперимент. Однако ее структура не является гибкой и не позволяет расширять описание фенотипа растений. Она также не позволяет учитывать параметры внешней среды. Интересная разработка – система PHENOME [10] – предлагает проводить фенотипирование растений в полевых условиях, используя карманный компьютер. Эта база данных позволяет собирать информацию о фенотипе томатов и хранить их в базе данных.

В настоящей работе для решения задачи сбора, интеграции, хранения и статистической обработки информации о растениях пшеницы мы предлагаем компьютерную систему WheatPGE (Wheat Phenotype, Genotype and Environment). Система хранит различные отношения, описывающие характеристики отдельного растения, и позволяет однозначно устанавливать взаимосвязь между генотипическими и фенотипическими признаками растений, а также параметрами окружающей среды. Применение системы позволит автоматизировать получение данных о взаимосвязи генотипа, фенотипа и окружающей среды у пшеницы, способствуя тем самым эффективному созданию новых сортов пшеницы с улучшенными свойствами.

2 Реализация

Для описания различных характеристик растений пшеницы нами была спроектирована реляционная база данных, которая лежит в основе системы

Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

WheatPGE и содержит более 23 таблиц, связанных между собой. В качестве сервера используется MySQL. Для работы с базой данных разработан веб-интерфейс, реализованный на основе модуля Catalyst – свободного кроссплатформенного программного каркаса, написанного на языке Perl. В Catalyst заложена методология разработки программного обеспечения MVC, в которой модель данных приложения, пользовательский интерфейс и управляющая логика разделены на три отдельных компонента. В результате модификация одного из компонентов оказывает минимальное воздействие на другие. Это позволяет добиться эффективной масштабируемости системы. Для связи базы данных с Catalyst используется технология ORM (объектно-реляционная проекция) – технология программирования, которая связывает базы данных с концепциями объектно-ориентированных языков программирования, создавая «виртуальную объектную базу данных». Технология позволяет связывать таблицы базы данных с объектами реального мира, например, объект генотип состоит из 9 связанных таблиц.

Важная особенность нашей системы – возможность для пользователя описывать произвольные морфологические признаки и параметры окружающей среды без помощи программиста. При этом происходит автоматическое расширение схемы базы данных, создается новая модель, описывающая объекты этого признака. Генерируются контроллеры и представления, реализующие базовые возможности работы с признаком (создание, удаление, редактирование). Этот подход имеет существенное ограничение. Семантическое описание нового признака ограничено одним реляционным отношением. Это означает, что описание должно укладываться в одну таблицу базы данных. Тем не менее, этого оказывается достаточным для описания большинства морфологических признаков и параметров окружающей среды, с которыми имеют дело экспериментаторы.

При занесении в базу большого количества гибридных генотипов становится актуальной задача визуализации схем скрещивания растений. Система WheatPGE позволяет автоматически визуализировать схемы скрещивания растений на основе информации об отношениях родитель – потомок, которые хранятся в базе данных. Схема представляется в виде ориентированного графа. Для размещения графа на плоскости и его рендеринга используется библиотека GraphViz (рис. 1).

Интерфейс системы WheatPGE реализован на основе сервера Apache с модулем mod_perl под управлением операционной системы CentOS Linux.

3 Структура базы данных WheatPGE

Центральным объектом базы данных является растение (рис. 2). Растение описывается как совокупность признаков генотипа, фенотипа и окружающей среды, в которой данное растение произрастает.

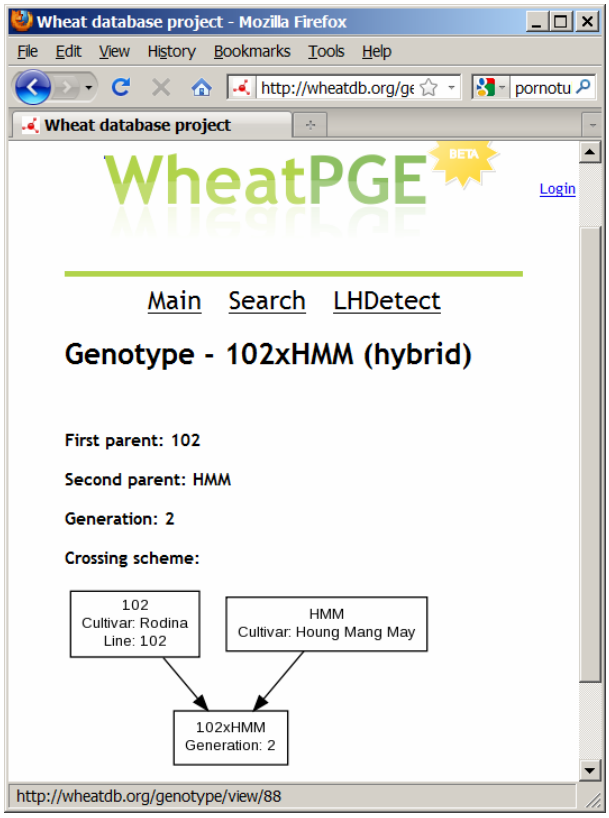

Рис. 1. Пример визуализации схемы скрещивания для гибридного генотипа

Рис. 2. Схема взаимосвязи между основными разделами информации в системе WheatPGE

3.1 Генотип

Описание генотипа растения содержит следующую информацию: сорт растения, линия (в случае, если растение из чистой линии) или родители (в случае, если растение – гибрид). Для родителей указываются ссылки на генотипы соответствующих растений. Дополнительно для гибридов можно указать поколение и материнское растение. Для генотипа можно определить список молекулярных маркеров (характеристик геномных ДНК, которые определяются экспериментально или могут быть импортированы из других баз данных). Маркеры объединяются в группы. Для каждого маркера из группы определяется его состояние (например, молекуляр-

ная масса или длина). Группа маркеров является характеристикой генотипа растения (рис. 3). При описании маркера указываются его тип, имя, список состояний и локализация на хромосоме.

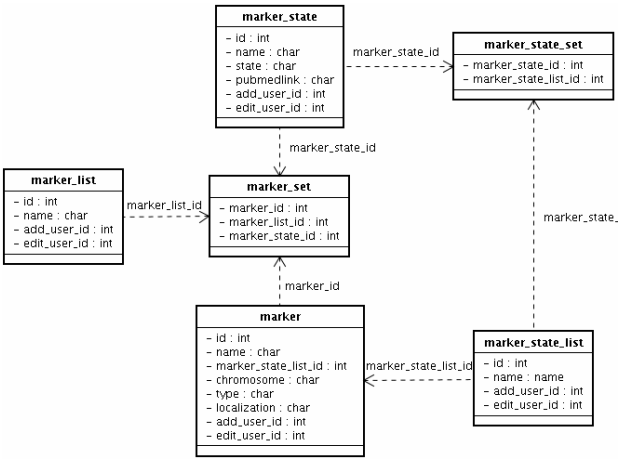

Рис. 3. Структурная схема реляционных отношений таблиц, описывающих молекулярные маркеры

3.2 Фенотип

Для описания фенотипа растения система WheatPGE позволяет создавать наборы отношений, каждое из которых содержит описание характеристик определенного морфологического признака (опушение листа, длина побега и колоса, количество колосьев, продуктивность и т. п.)

В текущей версии базы данных наиболее полно представлено описание такого морфологического признака, как опушение. Для него заданы следующие характеристики: плотность опушения (количество ворсинок (трихом) на единицу площади), вектор распределения трихом по длине. Система позволяет сохранять оцифрованные изображения морфологического признака, если это необходимо. Интерфейс для описания признака позволяет также подключать внешние программы анализа изображения для получения различных его характеристик, например, для получения информации о морфологических характеристиках опушения на основе анализа цифровых фотографий была использована программа LHDetect [4]. Структура базы данных позволяет легко расширять список анализируемых морфологических признаков растения и модифицировать информацию о них.

3.3 Окружающая среда

Подобно фенотипу WheatPGE позволяет расширять схему базы данных, добавляя произвольные параметры окружающей среды. Окружающая среда в базе данных может быть представлена набором таких характеристик, как место произрастания (теплица или открытый грунт), средние температура и количество осадков за сезон, дата посева семян и т. п.

3.4 Авторизация пользователей и разделение прав доступа

Пользователь может получить доступ к базе данных, зарегистрировавшись на сайте www.wheatdb.org. Зарегистрированный пользователь имеет возможность добавлять и аннотировать собственные растения. В каждой таблице базы данных содержатся поля, в которых прописываются идентификационный номер пользователя, создавшего запись в таблице, и идентификационный номер пользователя, отредактировавшего запись в таблице.

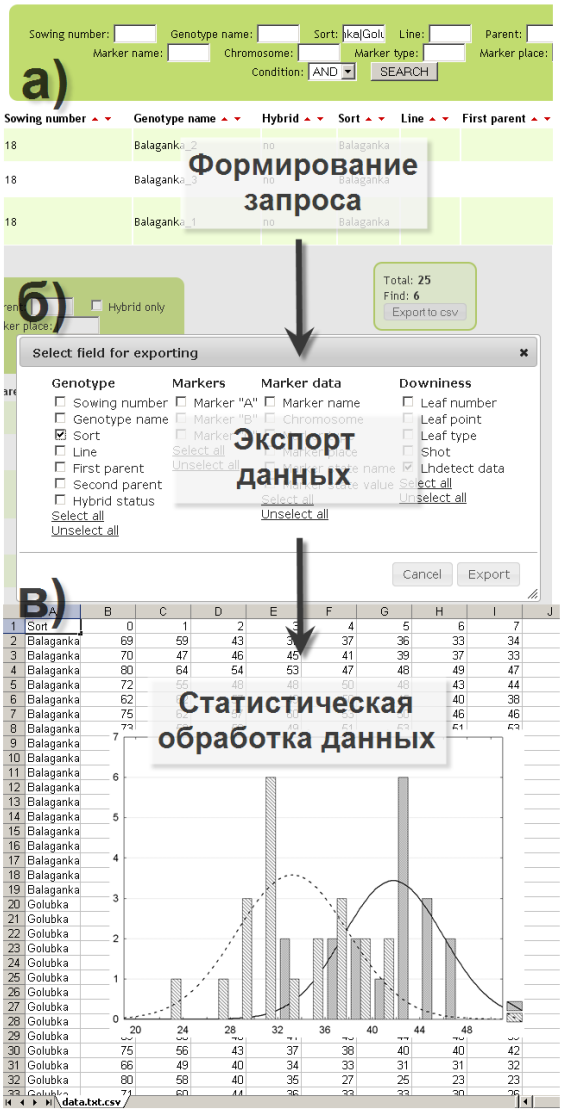

Рис. 3. Пример работы с системой WheatPGE. Извлечение данных о взаимосвязи сорта растения и его опушения: (а) интерфейс формирования запроса отбора растений; (б) выбор полей базы данных для экспорта в таблицу; (в) вид таблицы в Excel и статистический анализ распределения плотности опушения для сортов Балаганка и Голубка

Если для работы пользователю требуется возможность аннотировать дополнительные морфологические признаки или параметры окружающей среды, ему следует отправить запрос администрато-

ру системы с просьбой на расширение модели базы данных.

3.5 Пользовательский интерфейс

Пользователю предоставляется возможность просматривать списки генотипов, молекулярных маркеров, параметров окружающей среды и отдельные экземпляры растений, которые содержатся в базе. Кроме этого пользователь имеет возможность осуществлять поиск по растениям, которые содержатся в базе.

Поиск производится по следующим полям: посевной номер растения; название генотипа растения; сорт растения; линия; является ли растение гибридом или нет; название родительского генотипа; название молекулярных маркеров, которые присвоены генотипу растения; хромосома, на которой локализован молекулярный маркер; тип молекулярного маркера; положение молекулярного маркера на хромосоме.

При формировании запроса допустимо использование регулярных выражений, например, если необходимо найти в базе все растения двух сортов *Fora* и *Krasa*, в запросе достаточно написать *Fora*|*Krasa*.

Результаты любого запроса можно экспортировать в формате CSV с целью их дальнейшего анализа. При экспорте можно указать поля, необходимые для анализа. Экспортировать можно информацию о морфологических признаках растений, молекулярных маркерах и параметрах окружающей среды. Например, для анализа зависимости опушения листа от сорта растения пользователь должен на странице запроса указать список сортов растений (рис. 3а), которые он хотел бы включить в анализируемую выборку, и указать список характеристик опушения (рис. 3б). В итоге пользователь получает таблицу данных, в которой строкам соответствуют растения отобранных сортов, представленные в базе, а в колонках приводятся числовые характеристики опушения (рис. 3в). Такая таблица может быть далее проанализирована любой программой статистического анализа (Excel, Statistica и другие).

4 Выводы

В настоящее время база содержит более 250 аннотированных растений (более 100 сортов, более 1500 изображений листьев для анализа опушения).

Разработанная база данных позволяет устанавливать и анализировать взаимосвязь между генетическими и фенотипическими признаками растений и параметрами окружающей среды. Это обеспечивает решение целого ряда важных биологических задач. Например, исследование зависимости морфологических характеристик опушения листа от сорта растения, места произрастания, поиск генетических маркеров, статистически связанных с тем или иным типом опушения пшеницы и т. п.

Литература

[1] WheatPGE – system for analysis of relationships between genotype, phenotype end environment in wheat. – http://www.wheatdb.org/.

[2] 1001 Genomes Project –http://1001genomes.org/index.html.

[3] Дорошков А.В., Арсенина С.И., Пшеничникова Т.А., Афонников Д.А. Применение компьютерного анализа микроизображений листа для оценки характеристик опушения пшеницы Triticum aestivum L// Информационный вестник ВОГиС. – Новосибирск: Изд-во СО РАН, 2009. – Т. 13, № 1. – С. 218-226.

[4] Liying Zheng, Jingtao Zhang, Qianyu Wang: Mean-shift-based colour segmentation of images containing green vegetation// Computers and Electronics in Agriculture. – 2009. – V. 65. – P. 93-98.

[5] Bossu J., Géea Ch., Jones G., Truchetet F. Wavelet transform to discriminate between crop and weed in perspective agronomic images// Computers and Electronics in Agriculture. – 2009. – V. 65. – P. 133-143.

[6] Rodney M. Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology// Nature Reviews Genetics. – May 2001. – V. 2. – P. 370-381.

[7] GrainGenes – the international database for the wheat, barley, rye and oat genomes – http://www.graingenes.org.

[8] Ajjawi I., Lu Y., Savage L.J., Bell Sh.M., Last R.L. Large-scale reverse genetics in arabidopsis: case studies from the Chloroplast 2010 Project// Plant Physiology. – 2010. – V. 152. – P. 529-540.

[9] Exner V., Hirsch-Hoffmann M., Gruissem W., Hennig L. PlantDB – a versatile database for managing plant research// Plant Methods. – 2008. – V. 4, No 1.

[10] Vankadavath R.N., Hussain A.J., Bodanapu R., Kharshiing E., Basha P.O., Gupta S., Sreelakshmi1 Y., Sharma R. Computer aided data acquisition tool for high-throughput phenotyping of plant populations// Plant Methods. – 2009. – V. 5, No 18.

WheatPGE – system for analysis of the relationships between phenotype, genotype and environment in wheat

M.A. Genaev, A.V. Doroshkov, D.A. Afonnikov

We developed a WheatPGE system, the web-application for storing and processing of various morphological characteristics, genotype of the wheat plants and various environmental factors. The WheatPGE system allows analyzing the relationship between genetic and phenotypic traits of plants, as well as environmental conditions.

*Работа выполнена при финансовой поддержке интеграционных проектов СО РАН №№ 113, 26, 109 и Программы РАН «Происхождение и эволюция биосферы»