БИОФИЗИКА, 2020, том 65, № 2, с. 280–289

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 519.876.5

ГЕНЕТИЧЕСКИЕ ВАРИАНТЫ, АССОЦИИРОВАННЫЕ С ПРОДУКТИВНОСТЬЮ И СОДЕРЖАНИЕМ БЕЛКА И МАСЛА У СОИ

© 2020 г. А.А. Канапин*, **, А.Б. Соколкова*, А.А. Самсонова*, **, А.В. Щегольков*, С.В. Болдырев****, А.Ф. Аюпова****, Ф.Е. Хайтович****, С.В. Нуждин*****, М.Г. Самсонова***

**Санкт-Петербургский политехнический университет Петра Великого, 195251, С.-Петербург, Политехническая ул., 29*

***Санкт–Петербургский университет, 199034, Санкт-Петербург, Университетская наб., 7–9*

****Компания «Соевый комплекс», 350038, Краснодар, ул. Филатова, 19/2*

*****Сколковский институт науки и технологий, 143026, Москва, ул. Нобеля, 3*

******Университет Южной Калифорнии, CA 90089, Лос-Анджелес, США*

E-mail: m.samsonova @spbstu.ru

Поступила в редакцию 20.12.2019 г.
После доработки 20.12.2019 г.
Принята к публикации 22.01.2020 г.

Подходы, основанные на учете биоразнообразия, находятся на переднем крае создания новых сортов в соеводстве. В этой статье с помощью полногеномного поиска ассоциаций проведен анализ естественной вариабельности в популяции сортов культурной сои, используемых в отечественной селекции. Идентифицированы районы генома, контролирующие содержание белка и масла в семенах, а также продуктивность, из которых многие не были описаны ранее. Полученные результаты открывают путь к созданию новых маркеров для маркерной селекции этой культуры.

Ключевые слова: соя, полногеномный поиск ассоциаций, продуктивность, содержание балка и масла в семенах.

DOI: 10.31857/S000630292002009X

Соя (*Glycine max* (L.) Merr.) – один из самых динамичных и перспективных в России агрорынков, биологический потенциал которого сейчас почти не ограничен. Одним из ключевых требований времени к селекции сортов сои является проблема качества – содержание белка и масла в семенах, а также повышение продуктивности. Современные технологии молекулярного маркирования и полногеномного поиска ассоциаций могут значительно ускорить этот процесс.

Эти технологии используют результаты полногеномного секвенирования или генотипирования путем секвенирования (genotyping by sequencing, GBS). Генотипирование путем секвенирования – сравнительно дешевый метод, который «прочитывает» многочисленные короткие участки генома, позволяет выявить полиморфные позиции и тем самым охарактеризовать генетическое разнообразие и структуру популяции изучаемого вида [1]. Другим типом данных, необходимых для полногеномного поиска ассоциаций, являются данные фенотипирования. Технология полногеномного поиска ассоциаций делает возможной идентификацию локусов количественных признаков, контролирующих фенотипическую изменчивость по агрономически важным признакам, на основе которых затем могут быть разработаны маркеры для маркерной селекции [2]. Полногеномный поиск ассоциаций и маркерная селекция особенно эффективны в случае локусов большого эффекта, контролирующих просто наследуемые количественные признаки. В случае количественных, сложных признаков, контролируемых большим числом генов малого эффекта и зависящих от внешней среды, разработаны статистические методы предсказания селекционной ценности растений с использованием всей информации об изменчивости генома.

Селекция на качество у бобовых усложняется взаимодействием признаков, зависимостью их от внешней среды и значительным взаимодействием генотип х среда [3]. Значительные отрицательные корреляции были обнаружены между урожайностью и содержанием белка в семенах сои, а также между содержанием белка и масла, в то вре-

Сокращения: ОНП – однонуклеотидные полиморфизмы, QTL – локусы количественных признаков.

мя как урожайность и масло коррелируют положительно [4]. Было также показано, что условия культивирования влияют на эти признаки [5].

За истекшее двадцатилетие обнаружено много участков генома, контролирующих продуктивность и содержание белка и масла в семенах [3]. Однако из-за их частого плейотропного действия на отрицательно коррелированные признаки, а также из-за отсутствия у них большого эффекта и стабильности очень немногие из этих участков были далее использованы или включены в селекционные программы.

Генетическая основа современных сортов сужена многолетней селекцией, и одним из подходов к увеличению разнообразия, столь важного для селекции, является интрогрессия материала диких образцов или староместных сортов, широко используемых до начала «зеленой» революции. Неоценимым источником такой информации являются образцы коллекции ВИР (Всероссийского института генетических ресурсов растений им. Н.И. Вавилова) – самой большой в Европе и содержащей уникальный материал из всех соесеющих районов мира и для всех направлений использования этой высокобелковой культуры: пищевого, кормового и технического.

В этой статье методы геномики и биоинформатики были применены для описания генетического разнообразия образцов сои коллекции ВИР и современных культурных сортов, а также для идентификации участков генома, ассоциированных с хозяйственно важными признаками продуктивности – количеством белка и масла.

МАТЕРИАЛЫ И МЕТОДЫ

Материал. Исследуемая выборка из 280 сортов состояла из 121 образца сои коллекции ВИР (114 сортов *G. max*, 2 образца *G. gracilis* и 5 образцов мутантов и гибридов *G. soja*) и 160 образцов сои *G. max* коллекции компании «СоКо» (79 современных сортов и 80 линий из предварительного сортоиспытания).

Выращивание. Посев сортов и линий сои был произведен в Центральной зоне Краснодарского края 1 мая. Расстояние между рядами 70 см, расстояние между семенами 3 см. Почвенный покров экспериментального участка представлен выщелоченным слабогумусным сверхмощным тяжелосуглинистым черноземом. В целом почва обладает благоприятными водно-физическими свойствами и химическим составом для выращивания всех сельскохозяйственных культур, в том числе сои. Предшественник – озимая пшеница. Агротехника сои на экспериментально участке – рекомендованная для данной зоны выращивания. Посев осуществлялся механизированно с помощью селекционной кассетной сеялки. Всего опыт включал 191 четырехрядную делянку площадью 14 м^2 каждая. Все сортообразцы высеяны в двухкратной повторности. Кроме того, 139 сортообразцов, полученных от Всероссийского института генетических ресурсов им. Н.И. Вавилова, с ограниченным числом семян были высеяны на однорядных делянках площадью 3.5 м^2 без повторностей. Общая площадь участка под опытом составила 0.8 га. В период вегетации сои были проведены визуальные наблюдения за растениями, заключающиеся в фиксировании дат появления всходов и полного созревания для определения продолжительности вегетационных периодов исследуемых сортообразцов. Уборочные работы на экспериментальном опытном участке были проведены при полном созревании растений. Основаниями для назначения сроков уборки являются: опадение листьев, подсыхание вегетативных частей растения и снижение влажности семян до 14%. На четырехрядных делянках уборка проведена путем прямого комбайнирования с использованием селекционного комбайна. Скашиванию комбайном подвергали два средних (учетных) ряда четырехрядной делянки. Боковые рядки делянки не убирали – они являются защитными, так как испытывают влияние соседних делянок. Растения с однорядных делянок срезали вручную (серпом), формировали в сноп и затем обмолачивали с помощью селекционного комбайна.

Фенотипирование растений. Полученные семена были очищены от сорной примеси и взвешены с определением влажности. Продуктивность сортов (г/м^2) устанавливали путем деления массы семян с делянки при пересчете на стандартную (14%-ю) влажность на учетную площадь. Для определения биохимического состава семян (содержание белка и масла в процентах) использовали спектрометр ближней инфракрасной области, анализы на котором осуществляли в соответствии с ГОСТ Р 32749-2014. Анализ проводили на целых (неразрушенных) семенах, которые помещали в прибор в стандартных кюветах. Масса анализируемой навески семян составляла 8–10 г (40–60 шт.), время измерения одного образца – 30 с. Для получения достоверного результата были выполнены три параллельных измерения одного образца.

Генотипирование образцов. Геномная ДНК была разрезана двумя рестриктазами – *HindIII* и *NlaIII*. Использовали два типа адаптеров – баркоды, которые пришивали к концам, образованным разрезанием *HindIII,* и второй общий адаптер, который пришивали к свесам, образовавшимся при действии *NlaIII*. Количество полученных библиотек оценивали при помощи флуориметра Qubit (Thermo Fisher Scientific, США) и высокочувствительного набора для оценки концентра-

ции ДНК (Qubit DNA HS Assay Kit). Качество полученных библиотек оценивали на биоанализаторе Agilent 2100 (Agilent Technologies, Inc., США) с использованием высокочувствительного набора Agilent High Sensitivity DNA Kit. Секвенирование образцов, подготовленных по методу генотипирования путем секвенирования, проводили на секвенаторе HiSeq400 (Illumina, США) со следующими установками: длина прочтения – 150 нуклеотидов, парные чтения, длина индексного чтения – 7 нуклеотидов. По завершении запуска проводили конверсию в формат fastq с помощью программы bcl2fastq. Оценку качества файлов fastq проводили с помощью программы AfterQC [6], версия 0.9.6. Прочтения Illumina выравнивали на референсный геном сои G.max Wm82.a2.v1 с использованием программы bowtie2, версия 2.3.4.3 и параметров, взятых по умолчанию [7]. Поиск однонуклеотидных полиморфизмов (ОНП) выполнен программой NGSEP [8]. Фильтрация ОНП выполнена стандартно с сохранением полиморфизмов, отвечающих условиям Mapping Quality (MQ) > 40. Дальнейшая фильтрация ОНП выполнена с использованием VCFtools [9] и заключалась в выборе вариантов, у которых частота минорной аллели была больше 1%, а представленность образцов была выше 85%. Результирующее количество ОНП составило 2385.

Анализ фенотипических данных. Тест Шапиро–Уилка на нормальность [10] был применен к количественным фенотипическим признакам. Коэффициенты корреляции Спирмена для признаков были рассчитаны с использованием функции «rcorr» из библиотеки «Hmisc» R [11].

Оценка величины неравновесного сцепления. Неравновесное сцепление оценивали, вычисляя квадрат коэффициента корреляции (r^2) между генотипами. VCFtools [9] был использован для расчета внутрихромосомных значений r^2 и значений r^2 между ОНП на разных хромосомах. Неравновесное сцепление оценивали путем построения графика внутрихромосомных значений r^2 относительно физического расстояния (т.п.н.) между маркерами в R [12]. В качестве критического значения r^2 была принята 95 перцентиль значений r^2 между ОНП на разных хромосомах после трансформации с использованием квадратного корня. Убывание неравновесного сцепления оценивали путем построения регрессионной линии, используя подход Хилла и Вайра [13]. Пересечение регрессионной линии внутрихромосомных значений r^2 с пороговым значением r^2 считалось оценкой величины неравновесного сцепления.

Анализ структуры популяции. Структуру популяции образцов анализировали методом главных компонент (библиотека в R SNPRelate [14]) и с помощью пакета программ на языке R LEA [15]. Во втором анализе выбор количества субпопуляций был основан на критерии кросс–энтропии. Этот критерий основан на прогнозировании доли замаскированных генотипов (завершение матрицы) и на методе перекрестной проверки. Меньшие значения критерия кросс-энтропии обычно означают лучшие запуски. Для каждого смоделированного значения K в диапазоне от 1 до 10 были выполнены десять независимых запусков. Было выбрано значение K, для которого кривая кросс-энтропии выходит на плато ($K = 6$). Индивидуальный образец с более чем 55%-й идентичностью к одной из субпопуляций классифицировали как принадлежащий этой субпопуляции.

Поиск ассоциаций. Полногеномный поиск ассоциаций был выполнен с использованием однолокусной линейной смешанной модели, реализованной в пакетах программ plink (версия 1.9) [16] и rMVP (версия 0.99.17) (https://cran.r-project.org/web/packages/rMVP/index.html). Анализ 2385 ОНП методом главных компонент показал, что первые пять из них объясняют 36% дисперсии всех маркеров. Однолокусная модель была реализована с использованием первых пяти компонент в качестве ковариант для всех фенотипических данных. Для обнаружения значимых ассоциаций между признаком и ОНП использовали значение частоты ложного обнаружения [17], равное 0.05. Для дальнейшего анализа были использованы только варианты, для которых статистически надежная ассоциация с фенотипическим признаком была подтверждена обоими программными пакетами. С помощью базы данных Legume information system [18] была произведена аннотация значимо ассоциированных ОНП.

РЕЗУЛЬТАТЫ

Фенотипический анализ образцов. По всей изученной выборке в среднем продуктивность равнялась 160.8 г/м^2, содержание белка – 41.3%, содержание масла 21.3%. Для образцов коллекции ВИР средние значения этих фенотипических признаков 116.8 г/м^2, 41.9%, 20.1% соответственно, для образцов коллекции компании «СоКо» – 182.5 г/м^2, 40.9%, и 21.2%, а для линий из предварительного сортоиспытания этой компании – 205.1 г/м^2, 41.2% и 22.8%.

Признак продуктивности имеет среднюю положительную корреляционную связь с содержанием масла (коэффициент корреляции Спирмена, $r = 0.6$) и слабую отрицательную корреляционную связь с содержанием белка (коэффициент корреляции Спирмена, $r = –0.21$). Кроме того, признак содержание белка имеет среднюю отрицательную корреляционную связь с содержанием масла (коэффициент корреляции Спирмена, $r = –0.4$).

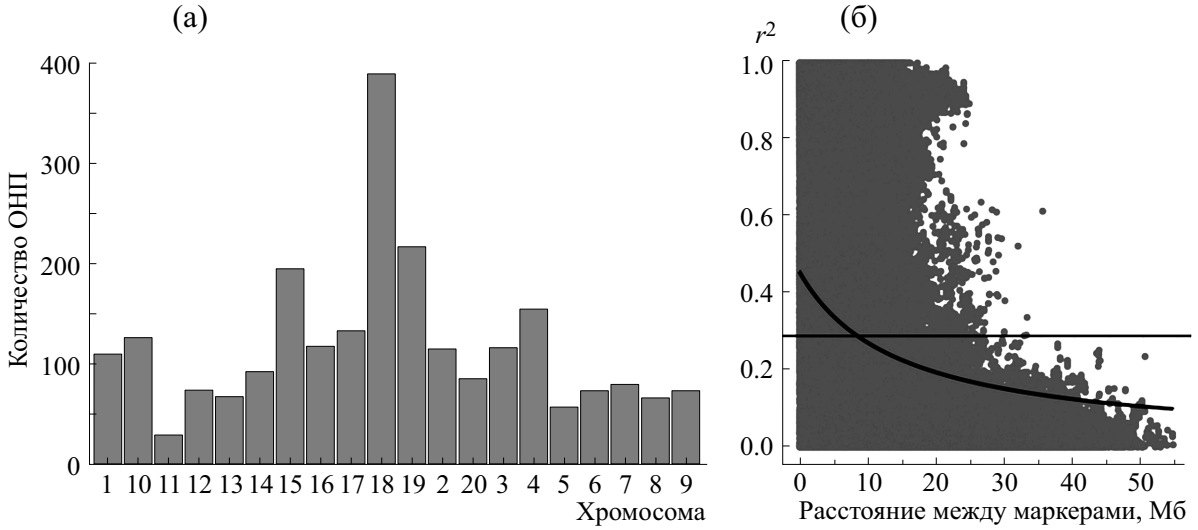

Рис. 1. (а) – Распределение снипов по 20 хромосомам генома сои. (б) – График неравновесного сцепления (r^2) для сои. Горизонтальная черная линия соответствует 95 перцентили значений r^2 между ОНП на разных хромосомах после трансформации с использованием квадратного корня.

Анализ полиморфизмов. Идентификацию однонуклеотидных вариантов в геноме образцов проводили методом генотипирования путем секвенирования. Идентифицированные ОНП были отфильтрованы для сохранения полиморфизмов, присутствующих по меньшей мере в 85% генотипов и имеющих частоту минорной аллели по меньшей мере 1%. Из рис. 1а видно, что результирующие ОНП распределены по всем 20 хромосомам сои. Наибольшее число ОНП находится в 18-й, 19-й и 15-й хромосомах, которые не являются самыми длинными по сравнению с остальными хромосомами.

Достаточность набора маркеров для полногеномного поиска ассоциаций прежде всего определяется величиной неравновесного сцепления. В меньших популяциях преимущественно самоопыляющихся организмов дрейф и отбор обычно имеют более сильные эффекты, чем рекомбинация, и, таким образом, неравновесное сцепление распространяется на большие области генома. График неравновесного сцепления между полиморфизмами образцов сои представлен на рис. 1б. В качестве критического значения было выбрано значение $r^2 = 0.29$. Регрессионная линия внутрихромосомных значений r^2 пересекает этот порог на физическом расстоянии примерно в 8.4 млн п.н. (рис. 1б).

Анализ структуры популяции. Характер дифференциации популяции был проанализирован методом главных компонент и визуализирован с помощью неукорененных филогенетических деревьев. На рис. 2а показан график, полученный методом главных компонент для первой и второй главных компонент, объясняющих изменчивость генетических данных. Как видно, образцы линий «CoKo» из предварительного сортоиспытания и образцы дикой сои и ее мутантов и гибридов образуют отдельные группы, а образцы коллекций ВИР и компании «CoKo» группируются вместе. В результате анализа в R-пакете LEA [15], в котором выбор количества субпопуляций основан на критерии кросс–энтропии, были выделены шесть субпопуляций (рис. 2б). При использовании 55% в качестве порога для отнесения образца к одной из субпопуляций 205 (62.5%) образцов были отнесены к одной из шести групп (табл. 1). Оставшиеся образцы являются результатом адмиксии, их не удалось однозначно отнести ни к одной из субпопуляций. К первой субпопуляции относятся образцы компании «CoKo», вторая субпопуляция преимущественно состоит из образцов коллекции ВИР и линий компании «CoKo». Третья и пятая субпопуляции, за исключением нескольких образцов полностью состоит из линий компании «CoKo». Четвертая и шестая субпопуляции преимущественно состоят из образцов коллекций ВИР и компании «CoKo», которые, как видно на рис. 2а, группируются вместе.

Анализ отдельных признаков. Полногеномный поиск ассоциаций был реализован с использованием первых пяти компонент в качестве ковариант для всех фенотипических данных в пакетах программ plink (версия 1.9) и rMVP (версия 0.99.17). Лучший тип анализа был выбран для каждого признака в отдельности на основе параметра геномного контроля (λ_{GC}).

Мы обнаружили 61 ОНП, значимо ассоциированный с содержанием масла в семенах, 63 ОНП, значимо ассоциированных с продуктивностью, и

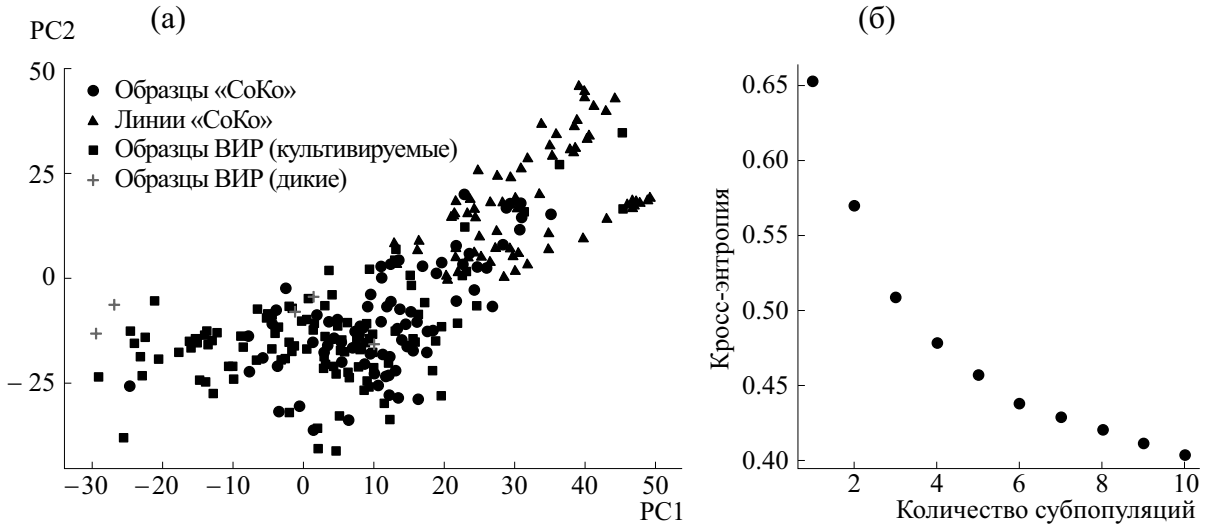

Рис. 2. (а) – График, полученный методом главных компонент для первой и второй главной компонент. (б) – График кросс-энтропии для образцов сои. Для каждого смоделированного значения K в диапазоне от 1 до 10 были выполнены десять независимых запусков. Было выбрано значение K, для которого кривая кросс-энтропии выходит на плато ($K = 6$).

35 ОНП, значимо ассоциированных с содержанием белка в семенах (рис. 3). Из них 25 ОНП имеют плейотропный эффект: 2 ОНП положительно и 13 ОНП отрицательно влияют на содержание масла в семенах и продуктивность, 10 ОНП положительно влияют на содержание белка и отрицательно на содержание масла в семенах, 4 ОНП, наоборот, – положительно на содержание масла и отрицательно на содержание белка в семенах. При этом 7 ОНП значимо ассоциированы со всеми тремя фенотипическими признаками.

Аннотация областей генома, где расположены значимые ОНП, показала, что внутри последовательностей известных генов находится 14, 5, 2 и 16 ОНП, значимо ассоциированных с содержанием масла в семенах, белка в семенах, масла и белка в семенах и с продуктивностью соответственно (табл. 2).

Из-за значительного неравновесного сцепления мы не можем однозначно идентифицировать причинно-следственные связи между этими ОНП и фенотипами. Тем не менее мы исследовали потенциальную природу ассоциированных генов (табл. 2). Так, например, ген Glyma.01g151300 содержит два ОНП, значимо ассоциированных с продуктивностью. Этот ген кодирует нодулин [19] и принадлежит суперсемейству вспомогательных белков (Major facilitator superfamily), осуществляющих транспорт широкого круга веществ через биомембраны [20]. ОНП Gm04:6140945, положительно влияющий на содержание масла в семенах и отрицательно – на содержание белка, находится внутри гена Glyma.04g073700, кодирующего амимоацил-тРНК-синтетазу. ОНП Gm15:10947679, ассоциированный с содержанием масла в семенах, находится в гене Glyma.15g135700, кодирующем ацетил-КоА-синтетазу, которая играет важную роль в деградации и синтезе липидов в семенах [21, 22]. Важную роль в биосинтезе жирных кислот играют 3-оксоацил [белок-переносчик ацил-группы] синтазы I, одну из которых кодирует ген Glyma.05g129600, в кото-

Таблица 1. Количество образцов в субпопуляциях

Образцы	Количество образцов в субпопуляциях						
	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	Admixsия
Образцы ВИР		29	1	20	4	17	58
Дикая соя и ее гибриды		4		1		1	3
Образцы компании «СоКо»	19	9	1	14	2	17	48
Линии компании «СоКо»		13	31		22		14

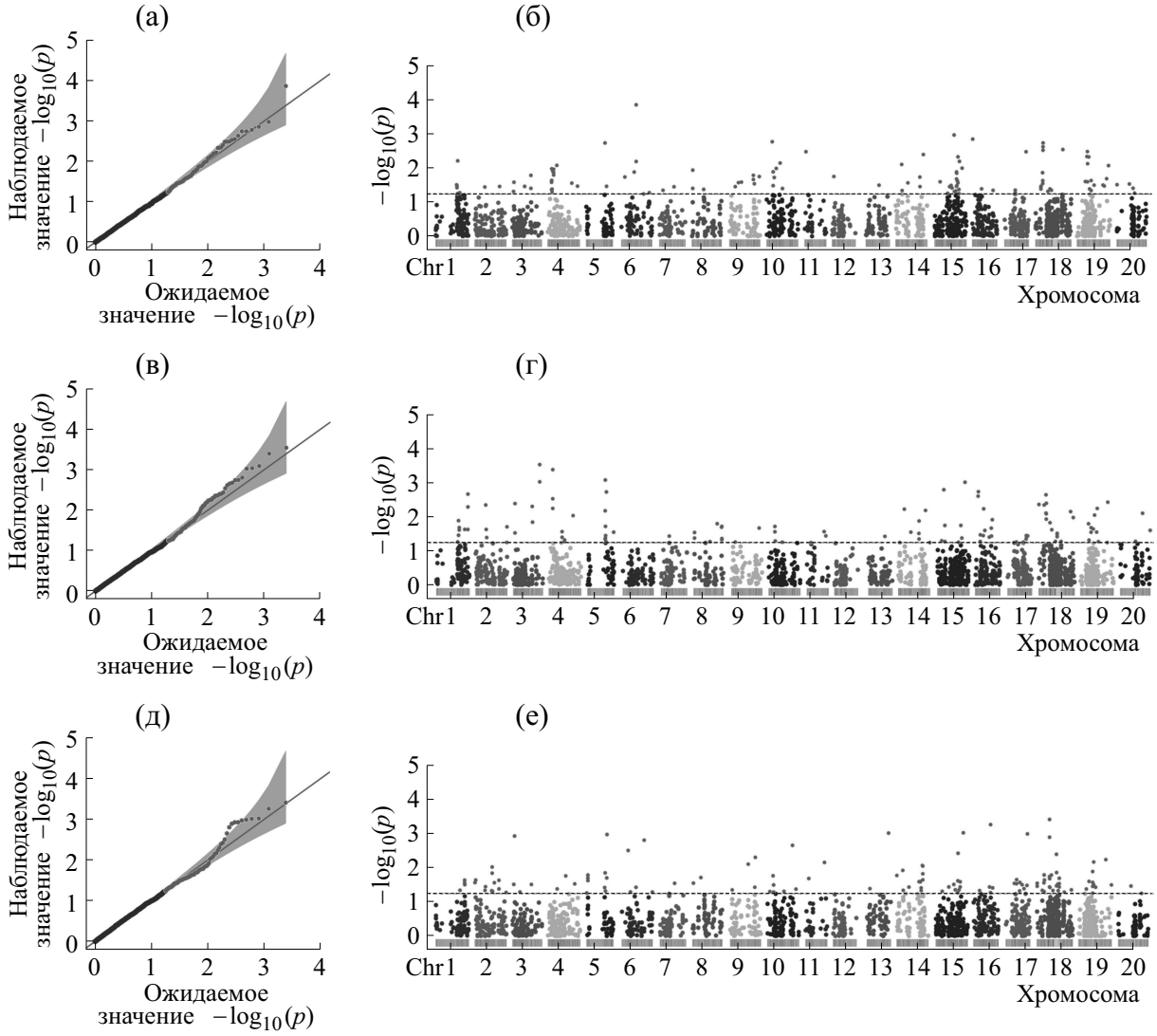

Рис. 3. (Результаты полногеномного поиска ассоциаций для продуктивности, содержания белка и масла в семенах. (а, в, д) – Графики QQ-plot для содержания масла, белка и для продуктивности. Эти графики показывают соответствие квантилей распределения значений q-статистики при справедливости нулевой гипотезы (отсутствие ассоциаций, нормальное распределение, *X*–ось) и квантилей распределения значений этой тестовой статистики, получаемым при тестировании (*Y*-ось). (б, г, е) – Манхеттановские графики для содержания масла, белка в семенах и продуктивности, *X*-ось – геномные координаты, *Y*–ось – отрицательный логарифм тестовой статистики ассоциации q для снипов, которые обозначены точками. Значимо ассоциированные полиморфизмы расположены выше пунктирной линии.

ром находится ОНП Gm05:32274272, ассоциированный с продуктивностью [23]. Ген Glyma.02g017500, содержащий ассоциированный с продуктивностью ОНП GM02:1553466, кодирует фотолиазу – фермент, участвующий в фоторепарации. Этот фермент вырезает из ДНК пиримидиновые димеры, возникновение которых индуцирует ультрафиолетовое излучение [24]. Ассоциированный также с продуктивностью ОНП Gm14:140838 локализуется в гене Glyma.14g001200, который кодирует кальцийсвязывающий белок с доменом EF-рука. Ортолог этого гена у *Glycine soja* участвует в ответе растения на солевой, осмотический и бикарбонатные стрессы [25]. ОНП Gm16:4345707, ассоциированный с содержанием белка в семенах, локализуется в гене Glyma.16g0045900, который кодирует синтаксин, белок, обеспечивающий специфичность слияния везикул при везикулярном транспорте [26]. ОНП Gm20:35399322, также ассоциированный с содержанием белка в семенах, локализуется в гене Glyma.20g111900, который кодирует транспортер катионов аминокислот [27]. У *A. thaliana* ген CAT9 транспортера катионов аминокислот 9 участвует в поддержании азот-зависимого гомеостаза аминокислот [28].

Таблица 2. Однонуклеотидные полиморфизмы внутри последовательностей известных генов

ОНП	Хромосома	Ген	Эффект	Признак
1_48811028_T_C	1	GLYMA_01G151300	+	
1_48811056_G_A			+	
2_449825_A_C	2	GLYMA_02G003800	–	Продуктивность
2_1553466_T_C		GLYMA_02G017500	–	
2_26061066_T_C		GLYMA_02G168900	+	
3_2161482_T_A	3	GLYMA_03G021000	+	Содержание масла
			+	Продуктивность
4_6140945_T_C	4	GLYMA_04G073700	+	Содержание масла
			–	Содержание масла
4_6163480_T_C		GLYMA_04G074000	+	Содержание масла
			–	Содержание масла
5_31849296_T_G	5	GLYMA_05G125100	–	Продуктивность
5_32274272_T_C		GLYMA_05G129600	–	
6_47133259_T_C	6	GLYMA_06G283400	+	
9_41467035_G_T	9	GLYMA_09G190100	+	Содержание масла
10_12734575_T_C	10	GLYMA_10G092000	–	
10_39381094_T_C		GLYMA_10G159500	–	Продуктивность
10_44728567_T_C		GLYMA_10G215100	+	
11_3245369_T_C	11	GLYMA_11G043800	–	Содержание масла
11_30226778_T_C		GLYMA_11G210100	+	Продуктивность
13_27118432_G_C	13	GLYMA_13G156200	+	Содержание масла
14_140838_G_A	14	GLYMA_14G001200	–	Продуктивность
15_10947679_C_T	15	GLYMA_15G135700	–	Содержание масла
16_659516_T_A	16	GLYMA_16G007900	+	
16_4345707_T_G		GLYMA_16G045900	–	Содержание масла
16_4782762_T_C		GLYMA_16G049900	–	
16_29296946_T_C		GLYMA_16G135500	+	Продуктивность
18_12210395_C_T	18	GLYMA_18G107800	–	Содержание масла
			–	Продуктивность
19_39885214_A_G	19	GLYMA_19G137500	–	
19_45429114_G_A		GLYMA_19G197100	–	Содержание масла
19_48356373_T_A		GLYMA_19G233800	–	
20_35399322_G_C	20	GLYMA_20G111900	–	Содержание масла

ОБСУЖДЕНИЕ

Подходы, основанные на учете биоразнообразия, находятся на переднем крае создания новых сортов в соеводстве. В этой статье с помощью полногеномного поиска ассоциаций проведен анализ естественной вариабельности в популяции сортов культурной сои, используемых в отечественной селекции, что в потенциале открывает путь к созданию новых маркеров для маркерной селекции этой культуры.

За истекшее двадцатилетие в многочисленных исследованиях обнаружено большое количество локусов количественных признаков (QTL), контролирующих содержание белка и масла в семенах и продуктивность, а также ОНП, ассоциированных с этими признаками [3]. Однако из-за отсутствия большого эффекта и стабильности, а также из–за негативной корреляции между содержанием белка и маслом /урожайностью и несогласованностью эффектов в разных условиях выращивания очень немногие из этих районов были далее использованы или включены в селекционные программы [29]. По данным Комитета по генетике сои (http://www.soybase.org), только два QTL, один на Хр. 15 (cqPro-15) и еще один на Хр. 20 (cqPro-20) обозначены как официально подтвержденные QTL, контролирующие содержание белка в семенах, и исследования подтвердили сегрегацию аллелей этих локусов во всех тестируемых популяциях (http://soybase.org/). QTL на Хр. 20 был в центре внимания нескольких исследований, в том числе по тестированию аллелей с высоким содержанием белка на разных генетических фонах, в этих экспериментах этот локус показал большой аддитивный эффект [5]. Важно, что в этом локусе обнаружены как аллели с отрицательной корреляцией между белком и продуктивностью, так и аллели с незначительным отрицательным влиянием на этот признак [30]. QTL на Хр. 20 также часто подтверждается при полногеномном поиске ассоциаций в различных популяций зародышевой плазмы [2, 4, 31], хотя уровень значимости этого района варьировал между исследованиями.

Данные о локусах, одновременно контролирующих продуктивность и качество семян, немногочисленны. Обнаружено, что локус E2, отвечающий за скорость созревания, обладает плейотропным эффектом и одновременно контролирует продуктивность и качество семян [32]. Контроль продуктивности также полигенный: в исследованиях по полногеномному поиску ассоциаций выявлены до 30 значимо ассоциированных ОНП, локализованных на 12 из 20 хромосом сои [33–35], некоторые из них воспроизводятся в разных условиях выращивания.

Не менее 110 QTL для содержания соевого масла были нанесены на карту, но использование этих QTL для селекции сои имеет некоторые ограничения [3]. Продемонстрирована стабильность в разных условиях выращивания QTL qOil-5-1, qOil-10-1 и qOil-14-1 [36]. Недавно авторы работы [37] провели сравнительный анализ нуклеотидных последовательностей между линиями с высоким и низким содержанием масла и выявили различие в ОНП и количестве копий генов, контролирующих биосинтез и деградацию липидов у этих линий. Сравнительный анализ показывает увеличенное количество копий гена белка транспорта липидов (LPT; Glyma.16g31780, Glyma.16g31840 и Glyma.16g31540) в линиях с высоким содержанием масла и большее число копий генов негативных регуляторов биосинтеза липидов (ABC-транспортер, Glyma.03g36310, Lipase3, и Glyma.13g04561) в линиях с низким содержанием масла. Возможно, это различие в числе копий генов обуславливает различие между линиями.

Наши результаты добавляют новое измерение в эти исследования благодаря включению в анализ сортов культурной сои, используемых в отечественной селекции, и позволяет выделить ассоциации с геномными областями, не обнаруженными в предыдущих анализах GWAS и QTL. Идентифицированные нами районы картируются в непосредственной близости от генов, участвующих в транспорте веществ через мембраны в клубеньках, синтезе белка и транспорте аминокислот, синтезе и деградации липидов в семенах, фоторепарации, реакции на стресс, а также генов, обеспечивающих специфичность слияния везикул при везикулярном транспорте. Некоторые из идентифицированных нами ОНП близки к уже известным ОНП, например ОНП Gm04:37264793, положительно влияющий на содержание масла в семенах и отрицательно – на содержание белка, локализуется на расстоянии 69 т.п.н. от ранее идентифицированного ОНП, ассоциированного с содержанием белка [38]. Ассоциированный с содержанием масла в семенах ОНП Gm19:45429114 локализован в районе длиной около 85 т.п.н., в котором находятся три ОНП, ассоциированных с продуктивностью [39] и числом семян [32]. Gm14:42540153, ассоциированный с содержанием масла в семенах, локализуется в районе длиной порядка 100 т.п.н., в котором идентифицированы четыре ОНП, ассоциированных с сухим весом 100 семян [40]. ОНП Gm06:51013713, Gm14:34210754 и Gm17:15598101, ассоциированные с продуктивностью, находятся соответственно на расстоянии 58 т.п.н., 45 т.п.н. и 56 т.п.н. от известных ОНП, ассоциированных с различными признаками, характеризующими урожайность (число семян в бобе, вес семян и сухой вес 100 семян соответственно) [32, 40, 41]. Наконец, ОНП Gm13:32199622, ассоциированный с содержанием белка, находится на расстоянии

26 т.п.н. от известного ОНП, ассоциированного с весом семян [39].

Результаты, полученные в этой статье, будут использованы для создания молекулярных маркеров с целью ускорения селекции сои и получения новых сортов.

БЛАГОДАРНОСТИ

Вычисления были проведены в Суперкомпьютером центре «Политехнический» СПбПУ и кластере Университета Южной Калифорнии. Исходные данные получены на базе уникальной научной установки Коллекция генетических ресурсов растений ВИР.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена в рамках и при финансовой поддержке Федеральной целевой программы (проект №14.575.21.0136 от 26.09.2017, уникальный идентификатор проекта RFMEFI57517X0136).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая работа не содержит описания каких-либо исследований с использованием людей и животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. C. B. Heim and J. D. Gillman, G3: Genes, Genomes, Genetics **7** (1), 299 (2017).
2. N. B. Bandillo, A. J. Lorenz, G. L. Graef, et al., Plant Genome **10** (2), 1 (2017). doi: 10.3835/plantgenome2016.06.0054
3. G. Kumawat, S. Gupta, M. B. Ratnaparkhe, et al., Front Plant Sci. **7**, 1852 (2016). doi: 10.3389/fpls.2016.01852
4. E.-U. Hwang, Q. Song, G. Jia, et al., BMC Genomics **15**, 1 (2014).
5. C. van Warrington, H. Abdel-Haleem, J. H. Orf, et al., Crop Sci. **54**, 963 (2014).
6. Sh. Chen, T. Huang, Y. Zhou, et al., BMC Bioinformatics **18**, 80 (2017).
7. B. Langmead and S. L. Salzberg, Nature Methods **9** (4), 357 (2012). doi: 10.1038/nmeth.1923.
8. D. Tello, J. Gil, C. D. Loaiza, et al., Bioinformatics, **35** (22), 4716 (2019). doi: 10.1093/bioinformatics/btz275
9. P. Danecek, A. Auton, G. Abecasis, et al., Bioinformatics **27**, 2156 (2011).
10. S. S. Shapiro and M. B. Wilk, Biometrika **52** (3–4), 591 (1965). doi: 10.1093/biomet/52.3-4.591
11. F. E. Harrell, Jr, *Hmisc: Harrell Miscellaneous. R package version 4.1-1.* URL: https://CRAN.R-project.org/package=Hmisc.
12. *The R Project for Statistical Computing.* URL: https://www.R-project.org/.
13. W. G. Hill and B. S. Weir, Theor. Popul. Biol. **33**, 54 (1988).
14. X. Zheng, D. Levine, J. Shen, et al., Bioinformatics **28** (24), 3326 (2012).
15. D. Falush, M. Stephens, and J. K. Mol. Ecol. Notes **7**, 574 (2007).
16. Ch. C. Chang, C. C. Chow, L. C. Tellier, et al., GigaScience **4**, 1 (2015).
17. J. D. Storey, Ann. Stat. **31**, 2013 (2003).
18. S. Dash, J. D. Campbell, E. K. Cannon, et al., Nucl. Acids Res. **44**, D1181 (2016).
19. D. P. S. Verma, M. G. Fortin, J. Stanley, et al., Plant Mol. Biol. **7**, 51 (1986).
20. E. M. Quistgaard, C. Löw, F. Guettou, and P. Nordlund, Nature Rev. Mol. Cell Biol. **17** (2), 123 (2016). DOI: 10.1038/nrm.2015.25
21. L. Yu, X. Tan, B. Jiang, Sun X, Gu S, Han T, et al., PLoS One **9** (7), e100144 (2014). DOI: 10.1371/journal.pone.0100144
22. T. R. Larson, T. Edgell, J. Byrne, et al., Plant J. **32** (4), 519 (2002). DOI: 10.1046/j.1365-313x.2002.01440.x
23. N. Li, Ch. Xu, Y. Li-Beisson, and K. Philippar, Trends Plant Sci. **21** (2), 145 (2016). DOI: 10.1016/j.tplants.2015.10.011
24. S. S. Gill, N. A. Anjum, R. Gill, et al., Sci. World J. **2015**, 250158 (2015). DOI: 10.1155/2015/250158
25. C. Chen, X. Sun, H. Duanmu, et al., PLoS One **10** (11), e0141888 (2015). DOI: 10.1371/journal.pone.0141888
26. F. Y. H. Teng, Y. Wang, and B. L. Tang, Genome Biol. **2**, reviews3012.1 (2001). DOI: 10.1186/gb-2001-2-11-reviews3012
27. W.-N. Fischer, D. D. F. Loo, et al., Plant J. **29** (6), 717 (2002).
28. H. Yang, Y.-D. Stierhof, and U. Ludewig, Front. Plant Sci. **6**, 212 (2015). DOI: 10.3389/fpls.2015.00212
29. J. Wang, P. Chen, D. Wang, et al., Mol. Breeding **35**, 92 (2015). DOI: 10.1007/s11032-015-0285-6
30. C. V. Warrington, H. Abdel-Haleem, D. L. Hyten, et al., Theor. Appl. Genet. **128** (5), 839 (2015). DOI: 10.1007/s00122-015-2474-4
31. H. Sonah, L. O'Donoughue, E. Cober, et al., Plant Biotechnol. J. **13** (2), 211 (2015). DOI: 10.1111/pbi.12249
32. C. Fang, Y. Ma, S. Wu, et al., Genome Biol. **18**, 161 (2017). doi: 10.1186/s13059-017-1289-9
33. B. W. Diers, J. Specht, K. M. Rainey, et al., G3: Genes, Genomes, Genetics **8** (10), 3367 (2018). DOI: 10.1534/g3.118.200332

34. A. Xavier, D. Jarquin, R. Howard, et al., G3: Genes, Genomes, Genetics **8** (2), 519 (2018). DOI: 10.1534/g3.117.300300
35. Ya. Jing, et al., Front. Plant Sci. **9**, 1392 (2018). DOI: 10.3389/fpls.2018.01392
36. Y. Cao, S. Li, Z. Wang, et al., Front. Plant Sci. **8**, 1222 (2017). DOI: 10.3389/fpls.2017.01222
37. B. Valliyodan, Dan Qiu, G. Patil, et al., Sci. Rep. **6**, 23598 (2016). DOI: 10.1038/srep23598
38. Y. Han, X. Zhao, D. Liu, et al., New Phytol. **209** (2), 871 (2016). DOI: 10.1111/nph.13626
39. R. I. Contreras-Soto, F. Mora, F. Lazzari, et al., Breed Sci. **67** (5), 435 (2017). DOI: 10.1270/jsbbs.17024
40. X. Li, X. Zhang, L. Zhu, et al., BMC Genetics **20** (1), 39 (2019). DOI: 10.1186/s12863-019-0737-9
41. Q. Song, D. L. Hyten, G. Jia, et al., G3: Genes, Genomes, Genetics **5** (10), 1999 (2015). DOI: 10.1534/g3.115.019000.

Genetic Variants Associated with Productivity and Protein and Oil Content in Soybeans

A.A. Kanapin*, **, A.B. Sokolkova*, A.A. Samsonova*, **, A.V. Schegolkov*, S.V. Boldyrev****, A.F. Aupova****, P.E. Khaitovich****, S.V. Nuzhdin*****, and M.G. Samsonova***

**Peter the Great St. Petersburg Polytechnic University, ul. Polytekhnicheskaya 29, St. Petersburg, 195251 Russia*

***St. Petersburg State University, Universitetskaya nab. 7/9, St. Petersburg, 199034Russia*

****The "SOKO" Company, ul. Filatova 19/2, Krasnodar, 350038 Russia*

*****Skolkovo Institute of Science and Technology, ul. Nobelya 3, Moscow, 143026 Russia*

******University of Southern California, CA 90089, Los Angeles, USA*

Biodiversity-based approaches are at the forefront of creating new varieties in soybean production. This paper presents an analysis of natural variability in the population of cultivated soybean varieties used in breeding in Russian Federation which was performed using Genome-Wide Association Studies. Genome regions controlling the protein and oil content in seeds as well as productivity have been identified, many of which have not been described previously. The obtained results open the way to the creation of new markers for marker selection of this crop.

Keywords: soybean, Genome-Wide Association Studies, productivity, protein and oil content in seeds