Вавиловский журнал генетики и селекции. 2021;25(1):57-63
DOI 10.18699/VJ21.007

Английский текст https://vavilov.elpub.ru/jour

Пангеномы сельскохозяйственных растений

А.Ю. Пронозин[1] ✉, М.К. Брагина[1, 2], Е.А. Салина[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр Института цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ pronozinartem95@gmail.com

Аннотация. Секвенирование генома организма – важный этап в его генетических исследованиях. Расшифровка геномной последовательности открывает широкие возможности для изучения строения структуры хромосом, распределения повторенных и кодирующих последовательностей, идентификации и аннотации генов. При исследовании сельскохозяйственных растений это позволяет анализировать функции генов, разрабатывать маркеры для поиска ассоциаций с фенотипическими признаками. При решении этих задач геном вида часто представлен последовательностью одного организма (так называемым референсным геномом). В последнее время, однако, появляется много свидетельств в пользу того, что большие структурные изменения генома, включая вариации числа копий генов и вариации наличия/отсутствия генов, преобладают в сельскохозяйственных культурах, играют ключевую роль в генетическом определении агрономически важных признаков и приводят к значительным вариациям функционального набора генов и генного состава у представителей одного вида. Такие структурные вариации не могут быть представлены на основе одной лишь референсной последовательности и описываются исходя из концепции пангенома. Пангеном – это информация о полном наборе генов таксона, среди которых можно выделить набор универсальных генов, общих для всех представителей таксона, и вариабельных генов, которые являются частично или полностью специфичными для его представителей. Анализ пангеномов дает более точное понимание генетического разнообразия генофонда. Технологии секвенирования и анализа пангеномов позволяют обеспечить возможность масштабного изучения геномных вариаций, доступ к более широкому спектру геномных данных в селекционных программах и помогут ускорить селекцию культурных растений для создания сортов со стабильно высокой урожайностью и устойчивостью к стрессам. В работе представлен краткий обзор исследования пангеномов сельскохозяйственных растений, описаны их структурные особенности, методы и программы биоинформатического анализа пангеномных данных.
Ключевые слова: сельскохозяйственные растения; геномы; пангеномы; гены; эволюция; биоинформатический анализ; вычислительные конвейеры.

Для цитирования: Пронозин А.Ю., Брагина М.К., Салина Е.А. Пангеномы сельскохозяйственных растений. *Вавиловский журнал генетики и селекции*. 2021;25(1):57-63. DOI 10.18699/VJ21.007

Crop pangenomes

A.Yu. Pronozin[1] ✉, M.K. Bragina[1, 2], E.A. Salina[1, 2]

[1] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ pronozinartem95@gmail.com

Abstract. Progress in genome sequencing, assembly and analysis allows for a deeper study of agricultural plants' chromosome structures, gene identification and annotation. The published genomes of agricultural plants proved to be a valuable tool for studing gene functions and for marker-assisted and genomic selection. However, large structural genome changes, including gene copy number variations (CNVs) and gene presence/absence variations (PAVs), prevail in crops. These genomic variations play an important role in the functional set of genes and the gene composition in individuals of the same species and provide the genetic determination of the agronomically important crops properties. A high degree of genomic variation observed indicates that single reference genomes do not represent the diversity within a species, leading to the pangenome concept. The pangenome represents information about all genes in a taxon: those that are common to all taxon members and those that are variable and are partially or completely specific for particular individuals. Pangenome sequencing and analysis technologies provide a large-scale study of genomic variation and resources for an evolutionary research, functional genomics and crop breeding. This review provides an analysis of agricultural plants' pangenome studies. Pangenome structural features, methods and programs for bioinformatic analysis of pangenomic data are described.
Key words: agricultural plants; genomes; pangenomes; genes; evolution; bioinformatics analysis; computational pipelines.

For citation: Pronozin A.Yu., Bragina M.K., Salina E.A. Crop pangenomes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):57-63. DOI 10.18699/VJ21.007

Введение

Секвенирование генома организма – важный этап в генетических исследованиях генома. Расшифровка геномной последовательности открывает широкие возможности для исследования строения структуры хромосом, распределения повторенных и кодирующих последовательностей, идентификации и аннотации генов (Брагина и др., 2019). Информация о последовательностях геномов разных видов позволяет проводить сравнительный филогенетический анализ для изучения отношений между видами, их происхождения и особенностей эволюции (Marchant et al., 2016; Wendel et al., 2016). У сельскохозяйственных растений все это дает возможность оценить влияние генетической изменчивости на функцию генов, определить гены, ответственные за наиболее ценные признаки сельскохозяйственных культур (Schnable et al., 2009; Wing et al., 2018).

При решении этих задач геном вида представляется последовательностью одного организма (так называемый референсный геном). Первичная структура референсного генома улучшается в результате целого ряда последовательных экспериментальных и биоинформатических исследований, ее аннотация служит отправной точкой для генетиков, исследующих данную культуру. Количество секвенированных, собранных и аннотированных референсных геномов растений увеличивается с каждым годом (Брагина и др., 2019). В версии 48 базы данных Ensembl plants (сентябрь 2020 г.) содержится 93 собранных и аннотированных генома растений (Howe et al., 2020). На основе референсной геномной последовательности и повторного секвенирования геномных последовательностей представителей одного вида (как правило, с использованием технологии коротких прочтений) производятся анализ генетической изменчивости, изучение однонуклеотидных полиморфизмов (single-nucleotide polymorphisms, SNPs) и крупных структурных вариаций (structural variations, SVs) генома. Последний тип вариаций наиболее труден для идентификации на основе секвенирования короткими прочтениями, однако с созданием технологий третьего поколения, позволяющих читать последовательности ДНК длиной до сотен тысяч нуклеотидов (Li et al., 2018), идентификация больших структурных перестроек становится более доступной и надежной. Появляется больше свидетельств в пользу того, что структурные изменения, включая вариации числа копий генов (copy number variations, CNVs) и вариации присутствия/отсутствия генов (presence/absence variations, PAVs), преобладают в сельскохозяйственных культурах и приводят к значительным вариациям функционального набора генов и генного состава у особей одного вида (Springer et al., 2009; Hirsch et al., 2014; Li et al., 2014; Lu et al., 2015; Zhao Q. et al., 2018).

Геномы и пангеном

Для более эффективного анализа и описания разнообразия генного состава была предложена концепция пангенома (Tettelin et al., 2005). Пангеном – это информация о полной выборке генов в биологическом кластере (таксоне), например виде, среди которых можно выделить набор универсальных (основных) генов, общих для всех образцов, и набор уникальных (вариабельных) генов, частично общих или индивидуально специфичных (Tettelin et al., 2005). Исследования пангенома до настоящего времени были сосредоточены на поиске наличия или отсутствия генов у объектов для определения универсального или уникального набора генов.

Термин «пангеном» был изначально сформулирован в работе (Tettelin et al., 2005) для бактериальных видов *Streptococcus agalactiae*. На сегодняшний день существует несколько определений этого термина, которые базируются на двух концепциях: структурной и функциональной (Tranchant-Dubreuil et al., 2018). Структурная концепция рассматривает пангеном как совокупность всех геномных последовательностей таксона. В рамках этой концепции нуклеотидные последовательности геномов-представителей таксона (одного вида или рода) сравниваются между собой, и на этой основе определяется их общий уникальный (не избыточный) набор фрагментов ДНК одинаковой длины (100 п. н. или больше, в зависимости от вида). Эти последовательности и описывают структуру пангенома (Snipen et al., 2009; Alcaraz et al., 2010).

Вторая концепция основана на его функциональном представлении. В качестве функциональной компоненты рассматриваются все кодируемые в нем гены. В этом случае пангеном может быть описан как объединение всех генов для представителей определенного таксона (Plissonneau et al., 2018). Однако для большого количества родственных организмов такой набор является вырожденным, поскольку они содержат много генов с высоким уровнем сходства первичной структуры и, соответственно, функций. Исключить избыточность пангенома можно за счет объединения сходных последовательностей генов в функциональные семейства (Sun et al., 2016). При этом гены-представители одного функционального семейства в разных организмах рассматриваются с точки зрения функции как одна последовательность.

Что касается таксономической принадлежности организмов, которые формируют пангеном, то, как правило, их набор ограничивается отдельным видом. Однако некоторые исследователи используют более широкую трактовку пангенома. Например, в работе В.В. Тец (2003) пангеном рассматривается как полный набор генов живых организмов, вирусов и мобильных элементов.

Структурные особенности пангенома

Гены в пангеноме можно разделить на две группы по их представленности в разных организмах (Golicz et al., 2016). К первой группе относятся гены, которые встречаются у всех представителей таксона. Такая группа генов называется универсальным набором (англ. core gene set). Вторую группу составляют гены, имеющиеся у части представителей таксона. Эту группу генов называют необязательными (indispensable), второстепенными (accessory) или вариабельными генами. Среди генов второй группы особо выделяют уникальные, представленные лишь у одного индивида в таксоне гены. Универсальные и вариабельные гены отражают функциональную основу и разнообразие представителей вида соответственно.

С точки зрения эволюции, универсальные гены в большинстве случаев являются генами, которые выполняют жизненно важные функции и они, как правило, сохра-

няются в пределах вида. Напротив, вариабельные гены и их особая фракция, уникальные гены, вносят вклад в разнообразие видов, что позволяет им адаптироваться к различным условиям окружающей среды. Доля уникальных генов в пангеноме изученных культур варьирует от 8 до 61 % (Tao et al., 2019). Однако полученный размер уникального генома, вероятно, будет недооценен из-за неспособности современных стратегий и технологий определять все функциональные изменения в генах.

На основании последовательности одного генома невозможно определить, какие гены – общие для всех представителей вида, а какие – только для некоторых. Тем не менее для каждой новой последовательности существует возможность идентифицировать, к какой части пангенома она относится: к универсальной или вариабельной. Чем больше геномов-представителей таксона секвенировано, тем больше обнаруживается уникальных генов. Это приводит к росту размера пангенома при увеличении количества геномов. Однако для набора универсальных генов увеличение количества геномов вызывает обратный процесс: часть генов, которые являются универсальными, у новых представителей вида может отсутствовать. В результате размер пангенома – совокупности всех различных генов вида – увеличивается, а предполагаемый размер универсального набора генов, как правило, уменьшается (Golicz et al., 2016; Wang et al., 2018). Схематически эта зависимость показана на рис. 1. Каждая точка на этом графике соответствует оценке количества генов в пангеноме для набора из *k*-геномов (взятых случайным образом из полной выборки *N* исследуемых геномов). При этом с увеличением *k* оценка общего количества генов в пангеноме растет (сплошная красная линия), а количество уникальных генов уменьшается (синяя штриховая линия). Примеры зависимостей для реальных пангеномов можно найти на сайте https://pangp.zhaopage.com. Таким образом, на оценку размера пангенома и долю универсальных генов в нем существенно влияет размер выборки организмов.

На размер и долю уникальных генов пангенома, помимо количества секвенированных геномов, также влияют: 1) выбор образцов для анализа – объединение диких и культурных видов даст пангеном с более высокой долей уникальных генов, чем использование только культурных растений (Montenegro et al., 2017; Zhao Q. et al., 2018); 2) уровень плоидности, способ размножения, эффект «бутылочного горлышка» в процессе доместикации и др. Виды растений с более высоким уровнем плоидности и аутбридинга и сокращением разнообразия в результате доместикации, как правило, имеют большую долю уникальных генов (Tao et al., 2019).

Можно предположить, что добавление неограниченного количества новых геномов в пангеном приводит к его неограниченному росту. Однако исследования разнообразия генов у видов сельскохозяйственных культур показали, что для них количество идентифицированных уникальных генов имеет тенденцию к уменьшению по мере увеличения числа секвенированных образцов. Это позволяет считать, что при определенном количестве представителей таксона включение дополнительных геномов в пангеном уже не приведет к дальнейшему увеличению количества его генов. Такие пангеномы называют закрытыми. У томата

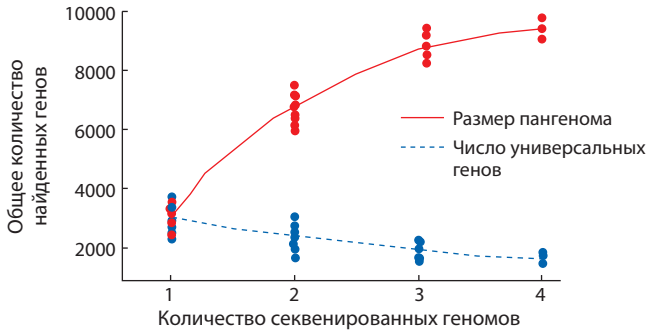

Рис. 1. Зависимость размера пангенома и числа универсальных генов в нем от числа секвенированных геномов-представителей таксона.

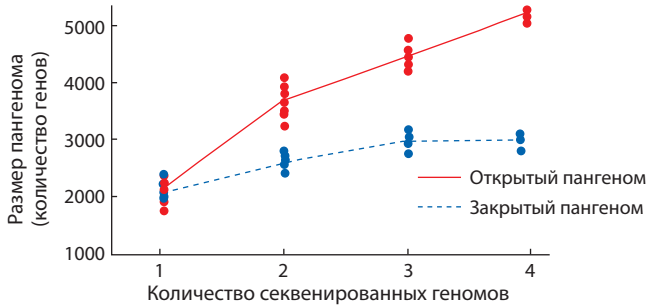

Рис. 2. Зависимость количества генов в пангеноме (ось *Y*) от количества секвенированных представителей таксона (ось *X*) для двух типов пангеномов: открытых и закрытых.

Для открытых геномов количество генов растет монотонно, для закрытых – выходит на плато.

(Gao et al., 2019), кукурузы (Hirsch et al., 2014), риса (Wang et al., 2018), сои (Li et al., 2014), подсолнечника (Hübner et al., 2019), *Brachypodium distachyon* (Gordon et al., 2017), *Brassica napus* (Hurgobin et al., 2018) и *B. oleracea* (Golicz et al., 2016) обнаружен закрытый пангеном.

Однако существуют также пангеномы, в которых общее количество генов растет при добавлении каждого нового образца. Такие пангеномы называют открытыми, они характерны для микроорганизмов. Например, результаты анализа пангенома грибного возбудителя септориоза листьев пшеницы *Zymoseptoria tritici* показали, что он относится к открытому типу (Plissonneau et al., 2018). Пангеном бактерии *Paenibacillus polymyxa*, обитающей в ризосфере растений и защищающей их от фитопатогенов (Zhou et al., 2020), также принадлежит к открытому типу.

При условии, что организмы из популяции отбираются случайным образом, тип пангенома можно оценить путем построения графика количества генов, обнаруженных в каждой новой геномной последовательности (рис. 2). Если после анализа определенного количества геномных последовательностей число генов в пангеноме выходит на плато, это считается характеристикой «закрытых» пангеномов. Такая зависимость схематически показана на рис. 2 (синяя штриховая линия). Если в зависимости размера пангенома от количества геномов нет признаков выхода на плато, – это характеристика «открытых» пангеномов. Зависимость числа генов от количества геномов для открытого пангенома схематически показана на рис. 2 (красная сплошная линия).

Сравнение размеров пангеномов и доли универсальной и вариабельной части для ряда растительных видов представлено в Приложении 1[1]. Данные в Приложении 1 демонстрируют, что количество представителей, включенных в анализ пангенома в растительных проектах, варьирует от 3 (репа, *Brassica rapa*) до 3000 (рис, *Oryza sativa*). Количество генов в пангеноме изменяется от 35 тыс. у риса – диплоида, до 128 тыс. у мягкой пшеницы – гексаплоида. Доля универсальных генов изменяется от 41 % у люцерны до 84 % у репы.

Функциональные особенности пангенома

По отношению функциональных особенностей генов из универсального и вариабельного наборов пангеномов исследования показывают, что универсальные гены отвечают за фундаментальные клеточные процессы, в то время как вариабельные гены ассоциированы, прежде всего, с функциями, которые могут дать преимущество в различных условиях окружающей среды. Так, при анализе пангенома трахинии двухколосковой *Brachypodium distachyon* (Gordon et al., 2017) было выявлено, что аннотации универсального набора генов обогащены такими терминами, как «гликолиз», «стероид», «гликозилирование», «кофермент». Аннотации генов вариабельного набора были более всего обогащены терминами «защитная функция», «развитие». В этой же работе показано, что отношение доли несинонимических замен к синонимическим у вариабельных генов выше, чем у универсальных. Кроме того, ортологи универсальных генов у риса и сорго оказались более консервативными, чем ортологи вариабельного набора генов. Универсальные гены также имеют более высокий уровень экспрессии, по сравнению с вариабельными (Gordon et al., 2017). Сходные результаты были получены при анализе пангенома сои *Glycine max* (Li et al., 2014; Liu et al., 2020), капусты (Golicz et al., 2016) и пшеницы (Montenegro et al., 2017).

Анализ этих и ряда других пангеномов сельскохозяйственных растений показал, что для них присуще следующее (Tao et al., 2019): последовательности вариабельных генов более изменчивы по сравнению с универсальными; скорость накопления несинонимических замен у вариабельных генов выше; вариабельные гены отличаются большим разнообразием функций; функциональные характеристики вариабельных и универсальных генов различаются, первые в большей степени связаны с ответом на факторы внешней среды, активностью рецепторов и передачей сигнала, вторые – с выполнением базовых клеточных функций. Таким образом, универсальные гены представляют собой консервативное ядро пангенома (и вида, соответственно), в то время как вариабельные гены – это мобильная его часть (как в качестве функций, так и в отношении первичной структуры и паттернов экспрессии).

Пангеномы и пантранскриптомы

Еще один из методов анализа генного состава у нескольких представителей какого-либо таксона – это анализ его транскриптомов. Нуклеотидные последовательности транскриптов (преимущественно мРНК), оценка уровня их экспрессии и наличие изоформ могут быть получены в результате высокопроизводительного секвенирования (RNA-seq), которое существенно дешевле, чем секвенирование генома. Транскриптомные данные позволяют оценить присутствие генов в геноме только в том случае, если они экспрессируются в какой-либо ткани или органе растения. Таким образом, по набору транскриптов нельзя представить полный состав генов в геноме, но получить приближенную оценку вполне возможно (особенно, если анализируется набор транскриптов из разных тканей на разных стадиях развития). При этом сборка транскриптома требует значительно меньше вычислительных ресурсов, а современные методы дают возможность получить ее с высоким качеством.

Исследование пантранскриптома 503 инбредных линий кукурузы дало возможность выявить генетическое разнообразие в белок-кодирующих генах: обнаружено более полутора миллиона однонуклеотидных вариаций, найдены мутации, ассоциированные с признаками развития растений (время прохождения ряда фаз роста) (Hirsch et al., 2014).

M. Jin с коллегами (2016) также изучали пантранскриптом 368 инбредных линий кукурузы. Они обнаружили более двух тысяч последовательностей, которые не были представлены в референсном геноме кукурузы, среди них гены, ответственные за ответ на биотический стресс. Рассмотрены вариации, ассоциированные с уровнем экспрессии генов (eQTL). Результаты были спроецированы на метаболические сети, что позволило уточнить механизмы их функционирования.

В работе (Ma et al., 2019) проанализировано 288 экспериментов по секвенированию транскриптома ячменя. Среди собранных транскриптов около 30 % не показали сходства с референсным геномом. Данные исследования пантранскриптома показали, что гены устойчивости к патогенам более многочисленны в дикорастущем ячмене. Такие гены в процессе доместикации были подвержены более сильному давлению отбора по сравнению с генами в других видах.

Методы сборки пангенома

В биоинформатическом анализе пангенома можно выделить основные этапы:

1. Сборка последовательностей пангенома.
2. Выделение консервативных и вариабельных участков геномных последовательностей.
3. Идентификация/предсказание и функциональная аннотация генов.
4. Идентификация полиморфизмов.
5. Хранение, обеспечение быстрого доступа и визуализация пангеномных данных.

Для сборки пангеномов существуют стратегии: сборка-выравнивание; метагеномный подход; выравнивание-сборка (Golicz et al., 2016; Hurgobin, Edwards, 2017; Tranchant-Dubreuil et al., 2018).

Сборка-выравнивание. Метод основан на сборке *de novo* последовательностей каждого представителя таксона отдельно, с последующим выравниванием последовательностей между собой, а также относительно

[1] Приложения 1–3 см. по адресу:
http://www.bionet.nsc.ru/vogis/download/pict-2021-25/appx2.pdf

референсного генома, для того чтобы уменьшить избыточность и определить набор общих и вариабельных участков последовательностей. Для сборки генома разработано несколько программных пакетов: Velvet (Zerbino, Birney, 2008), SOAPdenovo (Xie et al., 2014), ALLPATHS (Butler et al., 2008) и MaSuRCA (Zimin et al., 2013). Такой подход требует много времени и вычислительных ресурсов. Стратегия сборки *de novo* использована для анализа пангенома культивируемой сои (Li et al., 2010), дикой сои (Li et al., 2014), риса (Wang et al., 2018), капусты (Golicz et al., 2016), люцерны (Zhou et al., 2020).

Метагеномный подход заключается в объединении всех секвенированных прочтений от разных представителей таксона в один пул и последующей сборке *de novo* контигов пангенома на основе этих данных. Затем каждый собранный контиг относится к определенному геному путем выравнивания исходных прочтений этого представителя на метагеномную сборку и последующего оценивания покрытий контигов. Метагеномный подход позволяет работать с результатами секвенирования с низким уровнем покрытия. Его применяли для анализа геномов риса (Yao et al., 2015), томата (Gao et al., 2019).

Выравнивание-сборка. Эта стратегия использует сборку одного полного генома (референсной последовательности) в качестве основы для сборки геномов остальных представителей таксона (guide assembly). Прочтения из одного представителя вида выравниваются относительно референсного генома, те прочтения, что не совпали, отсеиваются и собираются отдельно. Последовательность референсного генома дополняется новыми собранными последовательностями, далее образцы сравниваются с данным референсным геномом. Выравнивание-сборка дает возможность сократить время построения пангенома. В случае, если геномный фрагмент присутствует сразу у нескольких представителей таксона, его последовательность будет собрана лишь один раз, в то время как при независимой сборке *de novo* этот фрагмент будет собираться столько раз, сколько представителей таксона было исследовано. Такой подход применен при анализе пангенома подсолнечника (Hübner et al., 2019).

Следует также отметить, что в ряде работ исследователи не использовали сборку геномных последовательностей, а выравнивали короткие прочтения на референсный геном. Это позволяет оценить связь однонуклеотидного полиморфизма с фенотипическими характеристиками растений. Существуют также методы, которые на основе выравнивания коротких прочтений дают возможность оценить структурные перестройки, дупликации и потери генов (Zhao et al., 2013). Метод выравнивания использовали при анализе пантранскриптома кукурузы (Hirsch et al., 2014), оценке изменения количества копий генов при анализе пангенома картофеля (Żmieńko et al., 2014).

Методы аннотации и анализа пангенома

С помощью аннотации пангенома можно идентифицировать последовательности генов в геномах представителей таксона, на основе сравнения их последовательностей определить ортологичные гены, а также семейства универсальных и вариабельных генов. Для автоматической аннотации пангеномов разработан ряд программных пакетов, выполненных в виде вычислительных конвейеров. Они проводят основные этапы анализа пангеномных последовательностей и их аннотации. Ниже – краткое описание возможностей ряда таких программ.

Программа PGAP (Zhao Y. et al., 2012) осуществляет масштабный поиск генов, проводит функциональную аннотацию, обогащение кластеров ортологичных генов терминами онтологии, анализ эволюции видов, выполняет структурный анализ пангенома, идентификацию универсальной и вариабельной части пангенома. В обновленной версии этой программы, PGAP-X (Zhao Y. et al., 2018), дальнейшее развитие получили методы представления и визуализации результатов анализа пангеномов.

Пакет программ PpsPCP (Tahir Ul Qamar et al., 2019) разработан для идентификации вариаций наличия/отсутствия генов (PAVs) в пангеномах. Анализ основан на полногеномном сравнении последовательностей представителей таксона и референсного генома в несколько раундов с последовательной коррекцией как набора генов, так и участков их выравнивания в референсном геноме. В результате создается набор генов пангенома путем объединения последовательностей отдельных геномов с референсным геномом и их аннотации.

Программа BPGA (Chaudhari et al., 2019) реализует широкие возможности по анализу пангеномов: кластеризация генов на основе сходства последовательностей, анализ наличия/отсутствия ортологов, построение графика зависимости размеров пангенома и его универсальной части от количества геномов, реконструкция филогенетического дерева между представителями таксона, анализ метаболических путей и функциональной аннотации, оценка отклонений GC состава, расчет различных статистических характеристик пангенома и др.

Программа panX (Ding et al., 2018) направлена на идентификацию кластеров ортологичных генов. Для этого используются кластеризация на основе сравнения последовательностей, верификация и уточнение состава кластеров на базе анализа эволюционных расстояний и филогенетической реконструкции; программа оценивает ассоциацию между генным составом индивидуальных представителей таксона и их фенотипов.

Программа Pan4Draft (Veras et al., 2018) разработана для получения улучшенной аннотации пангеномов за счет добавления к ней информации о последовательностях незавершенных геномов (unfinished genomes). Это геномы, у которых аннотация и сборка до уровня хромосом не завершены, но их последовательности содержат фрагменты геномной ДНК и представляют ценную информацию о разнообразии геномов вида. Методы анализа ряда пангеномов растений описаны в Приложениях 2 и 3.

Перспективы использования пангеномных данных

В настоящее время исследования в направлении секвенирования и анализа пангеномов сельскохозяйственных растений активно продолжаются и дают возможность получить все больше сведений о генетических вариациях и новых генах.

Одна из фундаментальных задач в изучении пангеномов сельскохозяйственных растений – оценка генетического

разнообразия их культурных представителей, а также диких сородичей. Такой анализ позволяет установить происхождение и эволюцию культурных растений, оценить влияние процесса селекции на генетическую структуру сортов. Анализ пангеномов, таким образом, отвечает на ряд важных вопросов о закономерностях эволюции геномов на уровне вида, механизмах возникновения новых генов, разнообразии функций генов и их ассоциациях с фенотипическими признаками растений.

Важным направлением исследования пангеномов сельскохозяйственных растений являются секвенирование и анализ геномов их диких сородичей. Предполагают, что дикие сородичи культурных растений могут содержать пул генов, связанных с адаптацией организмов к условиям окружающей среды, ответом на биотический и биотический стрессы, т. е. гены, которые могли быть утеряны представителями культурных растений в результате искусственного отбора (эффект «бутылочного горлышка») (Гончаров, Кондратенко, 2008; Гончаров, 2013; Purugganan, 2019). Обнаруженные гены могут быть в дальнейшем использованы для создания новых генотипов, более устойчивых к патогенам, вредителям и абиотическому стрессу. Таким образом, изучение пангеномов сельскохозяйственных растений не только имеет фундаментальный аспект, но также важно с точки зрения практической селекции.

Заключение

Более точное понимание генетического разнообразия генофонда в сочетании с передовыми технологиями секвенирования и высокопроизводительным фенотипированием может облегчить анализ признаков для выявления полезных генетических мутаций, позволить программам селекции получить доступ к более широкому спектру генетических ресурсов, помочь отбору лучших стратегий в селекционных программах и ускорить селекцию культурных растений для создания сортов со стабильно высокой урожайностью в стрессовых условиях.

Пангеномные исследования предлагают гораздо более широкое понимание генетического разнообразия генофондов сельскохозяйственных культур, чем анализ по ресеквенированию геномов, и, таким образом, могут быть чрезвычайно полезны для улучшения культурных растений. Тем не менее знания, полученные с помощью пангеномных исследований, требуют интеграции с QTL/GWAS и исследованиями по ресеквенированию геномов для определения важных генов и аллелей, которые будут использоваться в эффективной стратегии селекции.

Список литературы / References

Брагина М.К., Афонников Д.А., Салина Е.А. Прогресс в секвенировании геномов растений – направления исследований. *Вавиловский журнал генетики и селекции*. 2019;23(1):38-48. DOI 10.18699/VJ19.459.
[Bragina M.K., Afonnikov D.A., Salina E.A. Progress in plant genome sequencing: research directions. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2019;23(1): 38-48. DOI 10.18699/VJ19.459. (in Russian)]

Гончаров Н.П. Доместикация растений. *Вавиловский журнал генетики и селекции*. 2013;17(4/2):884-899.
[Goncharov N.P. Plants domestication. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013; 17(4/2):884-899. 2013;17(4/2):884-899. (in Russian)]

Гончаров Н.П., Кондратенко Е.Я. Происхождение, доместикация и эволюция пшениц. *Информационный вестник ВОГиС*. 2008; 12(1-2):159-179.
[Goncharov N.P., Kondratenko E.Ja. Wheat origin, domestication and evolution. *Informatcionniy Vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeders.* 2008;12(1-2):159-179. (in Russian)]

Тец В.В. Пангеном. *Цитология*. 2003;45(5):526-531.
[Tets V.V. Pangenome. *Citoligiya = Cytology.* 2003;45(5):526-531. (in Russian)]

Alcaraz L.D., Moreno-Hagelsieb G., Eguiarte L.E., Souza V., Herrera-Estrella L., Olmedo G. Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. *BMC Genomics*. 2010;11(1):332.

Butler J., MacCallum I., Kleber M., Shlyakhter I.A., Belmonte M.K., Lander E.S., Nusbaum C., Jaffe D.B. ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* 2008;18(5): 810-820. DOI 10.1101/gr.7337908.

Chaudhari N.M., Gupta V.K., Dutta C. BPGA-an ultra-fast pan-genome analysis pipeline. *Sci. Rep*. 2019;6(1):1-10. DOI 10.1038/srep24373.

Ding W., Baumdicker F., Neher R.A. panX: pan-genome analysis and exploration. *Nucleic Acids Res*. 2018;46(1):e5-e5. DOI 10.1093/nar/gkx977.

Gao L., Gonda I., Sun H., Ma Q., Bao K., Tieman D.M., Thannhauser T.W., Burzynski-Chang E.A., Fish T.L., Stromberg K.A., Sacks G.L., Foolad M.R., Diez M.J., Blanca J., Canizares J., Xu Y., Knaap E., Huang S., Klee H.J., Giovannoni J.J., Fei Z. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet*. 2019;51(6). DOI 1044.10.1038/s41588-019-0410-2.

Golicz A.A., Batley J., Edwards D. Towards plant pangenomics. *Plant Biotechnol. J*. 2016;14(4):1099-1105. DOI 10.1111/pbi.12499.

Gordon S.P., Contreras-Moreira B., Woods D.P., Des Marais D.L., Burgess D., Shu S., Stritt C., Roulin A.C., Schackwitz W., Tyler L., Martin J., Lipzen A., Dochy N., Phillips J., Barry K., Geuten K., Budak H., Juenger T.E., Amasino R., Caicedo A.L., Goodstein D., Davidson P., Mur L.A.J., Figueroa M., Freeling M., Catalan P., Vogel J.P. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun*. 2017;8(1):2184. DOI 10.1038/s41467-017-02292-8.

Hirsch C.N., Foerster J.M., Johnson J.M., Sekhon R.S., Muttoni G., Vaillancourt B., Peñagaricano F., Lindquist E., Pedraza M., Barry K., Leon N., Kaeppler Sh.M., Buell R.C. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*. 2014;26(1):121-135. https://doi.org/10.1105/tpc.113.119982.

Howe K.L., Contreras-Moreira B., De Silva N., Maslen G., Akanni W., Allen J., Carbajo M. Ensembl Genomes 2020 – enabling non-vertebrate genomic research. *Nucleic Acids Res*. 2020;48(D1):D689-D695. DOI 10.1093/nar/gkz890.

Hübner S., Bercovich N., Todesco M., Mandel J.R., Odenheimer J., Ziegler E., Lee J.S., Baute G.J., Owens G.L., Grassa C.J., Ebert D.P., Ostevik K.L., Moyers B.T., Yakimowski S., Masalia R.R., Gao L., Ćalić I., Bowers J.E., Kane N.C., Swanevelder D.Z.H., Kubach T., Muños S., Langlade N.B., Burke J.M., Rieseberg L.H. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants*. 2019;5(1):54-69. DOI 10.1038/s41477-018-0329-0.

Hurgobin B., Edwards D. SNP discovery using a pangenome: has the single reference approach become obsolete. *Biology*. 2017;6(1):21. DOI 10.3390/biology6010021.

Hurgobin B., Golicz A.A., Bayer P.E., Chan C.K., Tirnaz S., Dolatabadian A., Schiessl S.V., Samans B., Montenegro J.D., Parkin I.A., Pires J.C. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus. Plant Biotechnol. J*. 2018;16(7):1265-1274. DOI 10.1111/pbi.12867.

Jin M., Liu H., He C., Fu J., Xiao Y., Wang Y., Xie W., Wang G., Yan J. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Sci. Rep*. 2016;6:18936. DOI 10.1038/srep18936.

Li C., Lin F., An D., Wang W., Huang R. Genome sequencing and assembly by long reads in plants. *Genes*. 2018;9(1):6. DOI 10.3390/genes9010006.

Li R., Zhu H., Ruan J., Qian W., Fang W., Shi Z., Li Y., Li Sh., Shan G., Kristiansen K., Li S., Yang H., Wang J., Wang J. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res*. 2010;20(2):265-272. DOI 10.1101/gr.097261.109.

Li Y.H., Zhou G., Ma J., Jiang W., Jin L.G., Zhang Z., Guo Y., Zhang J., Sui Y., Zheng L., Zhang S.S. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol*. 2014;32(10):1045. DOI 10.1038/nbt.2979.

Liu Y., Du H., Li P., Shen Y., Peng H., Liu S., Zhou G., Zhang H., Liu Z., Shi M., Huang X., Li Y., Zhang M., Wang Z., Zhu B., Han B., Liang C., Tian Z. Pan-genome of wild and cultivated soybeans. *Cell*. 2020;182(1):162-176. DOI 10.1016/j.cell.2020.05.023.

Lu F., Romay M.C., Glaubitz J.C., Bradbury P.J., Elshire R.J., Wang T., Li Y., Li Y., Semagn K., Zhang X., Hernandez A.G. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun*. 2015;6:6914. DOI 10.1038/ncomms7914.

Ma Y., Liu M., Stiller J., Liu Ch. A pan-transcriptome analysis shows that disease resistance genes have undergone more selection pressure during barley domestication. *BMC Genomics*. 2019;20(1):12. https://doi.org/10.1186/s12864-018-5357-7.

Marchant D.B., Soltis D.E., Soltis P.S. Genome evolution in plants. *eLS*. 2016;1-8. DOI 10.1002/9780470015902.a0026814.

Montenegro J.D., Golicz A.A., Bayer P.E., Hurgobin B., Lee H., Chan C.K., Visendi P., Lai K., Doležel J., Batley J., Edwards D. The pangenome of hexaploid bread wheat. *Plant J*. 2017;90(5):1007-1013. DOI 10.1111/tpj.13515.

Plissonneau C., Hartmann F.E., Croll D. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol*. 2018;16(1):5. DOI 10.1186/s12915-017-0457-4.

Purugganan M.D. Evolutionary insights into the nature of plant domestication. *Curr. Biol*. 2019;29(14):R705-R714. DOI 10.1016/j.cub.2019.05.053.

Schnable P.S., Ware D., Fulton R.S., Stein J.C., Wei F., Pasternak S., Minx P. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 2009;326(5956):1112-1115. DOI 10.1126/science.1178534.

Snipen L., Almøy T., Ussery D.W. Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics*. 2009;10(1):385. DOI 10.1186/1471-2164-10-385.

Springer N.M., Ying K., Fu Y., Ji T., Yeh C.T., Jia Y., Wu W., Richmond T., Kitzman J., Rosenbaum H., Iniguez A.L., Barbazuk W.B., Jeddeloh J.A., Nettleton D., Schnable P.S. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet*. 2009;5(11):e1000734. DOI d10.1371/journal.pgen.1000734.

Sun C., Hu Z., Zheng T., Lu K., Zhao Y., Wang W., Shi J., Wang C., Lu J., Zhang D., Li Z., Wei C. RPAN: rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Res*. 2016;45(2):597-605. DOI 10.1093/nar/gkw958.

Tahir Ul Qamar M., Zhu X., Xing F., Chen L.L. ppsPCP: a plant presence/absence variants scanner and pan-genome construction pipeline. *Bioinformatics*. 2019;35(20):4156-4158. DOI 10.1093/bioinformatics/btz168.

Tao Y., Zhao X., Mace E., Henry R., Jordan D. Exploring and exploiting pan-genomics for crop improvement. *Mol. Plant*. 2019;12(2):156-169. DOI 10.1016/j.molp.2018.12.016.

Tettelin H., Masignani V., Cieslewicz M.J., Donati C., Medini D., Ward N.L., Angiuoli S.V., Crabtree J., Jones A.L., Durkin A.S., DeBoy R.T. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. USA*. 2005;102(39):13950-13955. DOI 10.1073/pnas.0506758102.

Tranchant-Dubreuil C., Rouard M., Sabot F. Plant pangenome: impacts on phenotypes and evolution. *Ann. Plant Rev. Online*. 2018;453-478. DOI 10.1002/9781119312994.apr0664.

Veras A., Araujo F., Pinheiro K., Guimarães L., Azevedo V., Soares S., Costa da Silva A., Ramos R. Pan4Draft: a computational tool to improve the accuracy of pan-genomic analysis using draft genomes. *Sci. Rep*. 2018;8(1):1-8. DOI 10.1038/s41598-018-27800-8.

Wang W., Mauleon R., Hu Z., Chebotarov D., Tai S., Wu Z., Li M., Zheng T., Fuentes R.R., Zhang F., Mansueto L. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*. 2018;557(7703):43. DOI 10.1038/s41586-018-0063-9.

Wendel J.F., Jackson S.A., Meyers B.C., Wing R.A. Evolution of plant genome architecture. *Genome Biol*. 2016;17:37. DOI 10.1186/s13059-016-0908-1.

Wing R.A., Purugganan M.D., Zhang Q. The rice genome revolution: from an ancient grain to Green Super Rice. *Nat. Rev. Genet.* 2018;19:505-517. DOI 10.1038/s41576-018-0024-z.

Xie Y., Wu G., Tang J., Luo R., Patterson J., Liu S., Zhou X., Lam T., Li Y., Xu X., Wong G.K., Wang J. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*. 2014;30(12):1660-1666. DOI 10.1093/bioinformatics/btu077.

Yao W., Li G., Zhao H., Wang G., Lian X., Xie W. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.* 2015;16:187. DOI 10.1186/s13059-015-0757-3.

Zerbino D.R., Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821-829. DOI 10.1101/gr.074492.107.

Zhao M., Wang Q., Wang Q., Jia P., Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics.* 2013;14(1). DOI 10.1186/1471-2105-14-S11-S1.

Zhao Q., Feng Q., Lu H., Li Y., Wang A., Tian Q., Zhan Q., Lu Y., Zhang L., Huang T., Wang Y. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet*. 2018;50(2):278-284. DOI 10.1038/s41588-018-0041-z.

Zhao Y., Sun C., Zhao D., Zhang Y., You Y., Jia X., Yang J., Wang L., Wang J., Fu H., Kang Y., Chen F., Yu J., Wu J., Xiao J. PGAP-X: extension on pan-genome analysis pipeline. *BMC Genomics.* 2018;19(1):115-124. DOI 10.1186/s12864-017-4337-7.

Zhao Y., Wu J., Yang J., Sun S., Xiao J., Yu J. PGAP: pan-genomes analysis pipeline. *Bioinformatics*. 2012;28(3):416-418. DOI 10.1093/bioinformatics/btr655.

Zhou L., Zhang T., Tang S., Fu X., Yu Sh. Pan-genome analysis of *Paenibacillus polymyxa* strains reveals the mechanism of plant growth promotion and biocontrol. *Antonie van Leeuwenhoek.* 2020;113:1539-1558. DOI 10.1007/s10482-020-01461-y.

Zimin A.V., Marçais G., Puiu D., Roberts M., Salzberg S.L., Yorke J.A. The MaSuRCA genome assembler. *Bioinformatics*. 2013; 29(21):2669-2677. DOI 10.1093/bioinformatics/btt476.

Żmieńko A., Samelak A., Kozłowski P., Figlerowicz M. Copy number polymorphism in plant genomes. *Theor. Appl. Genet.* 2014;127:1-18. DOI 10.1007/s00122-013-2177-7.

ORCID ID

A.Yu. Pronozin orcid.org/0000-0002-3011-6288
E.A. Salina orcid.org/0000-0001-8590-847X

Благодарности. Работа выполнена при поддержке Российского научного фонда, грант № 18-14-00293.
Авторы благодарны Н.А. Шмакову и Д.А. Афонникову за помощь в работе над текстом статьи. Считаем своим приятным долгом поблагодарить анонимных рецензентов за ценные замечания.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 04.11.2020. После доработки 27.12.2020. Принята к публикации 03.01.2021.