Биотехнология. Теория и практика. 2013, №4, стр. 20-28
DOI: 10.11134/btp.4.2013.3

УДК 577.21

МОЛЕКУЛЯРНЫЕ МАРКЕРЫ. ПРИЧИНЫ И ПОСЛЕДСТВИЯ ОШИБОК ГЕНОТИПИРОВАНИЯ

М.Е. Омашева, К.П. Аубакирова, Н.А. Рябушкина

РГП «Институт биологии и биотехнологии растений», г. Алматы
natrya7@yahoo.com

Молекулярные маркеры, в основу которых положена реакция гибридизации или этап ПЦР, выявляющие полиморфизм ДНК, используются в настоящее время в различных областях биологии, в том числе в изучении и сохранении генетического разнообразия, идентификации индивидуумов, филогенетике, картировании полезных признаков качества и устойчивости к стрессовым факторам, в селекционном процессе, биотехнологии и др. До начала эксперимента исследователи должны определить, какой тип маркеров использовать исходя из следующих критериев: вариабельность и количество требуемых маркеров, необходимость в их кодоминантности, соответствующие требования к выделяемой ДНК; практические – эффективность, воспроизводимость анализа, необходимое соответствующее техническое обеспечение и стоимость. Для корректной интерпретации результатов генотипирования следует учитывать тот факт, что применение любого типа молекулярных маркеров сопряжено с рядом ошибок генотипирования, главные из них – выпадение бóльших аллелей, «нуль» аллели, «stutter» аллели вследствие особенностей Таг-полимеразы, негомологичность амплифицированных последовательностей одинакового размера (гомоплазия). Исследователи формулируют определяющие условия для уменьшения количества ошибок при генотипировании и снижения их влияния на конечный анализ. В их числе качество и количество анализируемой ДНК, уровень технических возможностей и профессионализма персонала, поскольку человеческий фактор определяется как одна из главных причин некорректных результатов; проведение пилотных экспериментов для сравнительной оценки теоретического и реального коэффициента ошибки. Минимизирование ошибок достигается оценкой возможностей того или иного типа и скринингом маркеров; оптимизацией экспериментальных методов, надлежащим использованием контролей, повторностей, а также разработкой статистических подходов для выявления ошибок. Компромисс между выбраковкой локусов, порождающих ошибки и повышением потенциала оставляемых локусов для усиления генетического сигнала может быть различным в различных исследованиях, но главное, чтобы этот сигнал не был потерян в угоду «приемлемого» уровня ошибок и исследователи разрабатывают эмпирические подходы для достижения желаемого компромисса.

Ключевые слова: молекулярные маркеры, генотипирование, ошибки генотипирования

РАЗНОВИДНОСТИ И ОСОБЕННОСТИ МОЛЕКУЛЯРНЫХ МАРКЕРОВ

Прогресс в современной биологии в значительной степени базируется на развитии и использовании молекулярно-генетических подходов, в основе которых лежит анализ полиморфизма ДНК, выявляемый с помощью различных типов молекулярных маркеров. Молекулярные маркеры используются в настоящее время для генотипирования при оценке генетического родства/разнообразия между индивидами, сортами, для создания генетических карт, картирования генов интереса, в селекционном процессе (MAS, marker assisted selection), популяционной генетике, филогенетических исследованиях, биотехнологии и др. [1-5].

Эволюция молекулярных маркеров идет в направлении повышения разрешающей способности, быстроты, простоты и, по возможности, снижения стоимости анализа. Исследователями [6] сформулированы требования к ДНК маркерам, согласно которым идеальный молекулярный маркер должен отвечать комплексу характеристик: быть высоко полиморфным для выявления генетического разнообразия; иметь кодоминантное наследование, позволяя определять гомозиготное и гетерозиготное состояния диплоидных организмов; маркер должен случайно и часто быть распределен по геному; его проявление должно быть нейтральным (поскольку ДНК последовательности любого организма нейтральны к внешним условиям и осуществляемым процедурам); метод должен быть прост и дешев в использовании; высоко воспроизводим; должен позволять обмен результатами между лабораториями. Наличие нескольких типов маркеров объясняется тем, что ни один из них с высокой степенью не соответствует всей сумме перечисленных критериев. При этом очевидно, что особенности той или иной разновидности молекулярных маркеров обуславливают их более успешное применение для решения тех или иных задач. Исследователи перед началом эксперимента должны определить, какой тип маркеров использовать исходя из следующих критериев: вариабельность и количество требуемых маркеров, необходимость в их кодоминантности, соответствующие требования к растительному

материалу и выделяемой ДНК, размер генома и плоидность таксона; практические – эффективность анализа, стоимость и соответствующее техническое обеспечение [7]. Более того, тенденцией современных исследований с привлечением молекулярных маркеров является использование двух и более типов маркеров (описание типов см. ниже), что приводит к более однозначным, достоверным результатам. Так, для понимания механизмов устойчивости *V. amurensis* к *Agrobacterium* с помощью RAPD маркеров в картирующей популяции винограда был выявлен локус устойчивости, на основе которого были разработаны сопряженные с устойчивостью SCAR маркеры. Однако, точную локализацию этих маркеров удалось выяснить с помощью SSR маркеров с использованием референсной карты [8].

В последние десятилетия разработаны типы молекулярных маркеров, в основу которых положена реакция гибридизации, таких как RFLP, и несколько типов, включающих этап ПЦР (RAPD, AFLP, SSR, SNP, CAPS, SCAR, маркеры на основе ретротранспозонов и др.). В обзоре [6] приведен список наиболее часто используемых молекулярных маркеров, приложения их использования, преимущества и недостатки, а также сравнительная оценка, основанная на ряде признаков. В их числе: требования к количеству и качеству ДНК, количество анализируемых полиморфных локусов, простота использования, возможность автоматизации процесса, воспроизводимость и стоимость.

Ниже приведены наиболее часто используемые типы молекулярных маркеров.

RFLP (restriction fragment length polymorphism; полиморфизм длины рестрикционных фрагментов) – первый метод молекулярных маркеров для профилирования ДНК. Метод основан на технике расщепления с помощью особых рестрикционных ферментов (эндонуклеаз рестрикции) молекул ДНК, различающихся в гомологичных участках и соответственно мест рестрикции, и сравнении длин полученных фрагментов различных видов и даже линий живых организмов. Полученные фрагменты разделяются электрофоретически. После переноса на мембрану они идентифицируются гибридизацией с радиоактивно мечеными пробами (Southern blotting). Гибридизация позволяет определять длины фрагментов, комплементарных пробам. Каждый фрагмент рассматривается как аллель и используется в генетическом анализе. Достаточная степень полиморфизма, кодоминантность, частота и равномерное распределение по геному, высокая воспроизводимость позволили использовать метод в ДНК профилировании, картировании генома, локализации генов, имеющих отношение к болезням, генотипировании, изучении филогенеза. Согласно PubMed данным (plant+RFLP), максимальное количество публикаций с использованием этих маркеров приходится на 2003-2005 гг. Определенные недостатки метода – необходимость больших количеств ДНК высокого качества, использование изотопов, продолжительность анализа и др. Соответствующее развитие более эффективных и дешевых типов маркеров к настоящему моменту существенно снизили использование RFLP анализа.

Маркеры с использованием полимеразной цепной реакции (PCR)

PCR (polymerase chain reaction; ПЦР, полимеразная цепная реакция*)* – использование термостабильной ДНК полимеразы для амплификации *in vitro* отдельных/специфических последовательностей или локусов ДНК с применением случайных или специфических праймеров (олигонуклеотидных последовательностей). Амплифицированные фрагменты разделяются электрофоретически и полосы (пики) выявляются окрашиванием или радиоавтографией. Предпочтение методов генотипирования с использованием ПЦР обусловлено, прежде всего, потребностью в реакции небольших количеств ДНК (5-100 нг образца на реакцию).

Открытие и разработка полимеразной цепной реакции создали возможности для дизайна широкого спектра молекулярных маркеров. Развитием RFLP анализа явился метод ***CAPS*** (cleaved amplified polymorphic sequence) – амплификация в ПЦР последовательностей, вырезаемых по месту рестрикции соответствующих узнаванию рестриктазой нуклеотидных последовательностей полиморфных гомологичных участков. Различия проявляются в легко различимых по длине продуктов - фрагментов ДНК при электрофорезе. CAPS маркеры являются кодоминантными и позволяют выявлять полиморфизм в большом количестве индивидуумов. Например, разработка CAPS маркеров на основе последовательностей определенных генов и EST последовательностей позволила исследователям создать соответствующие пары праймеров, с помощью которых была установлена подлинность 67 сортов чая (95%), культивируемых в Японии [9]. Данный тип маркеров помогает выявлению недавней эволюционной истории, пониманию направления эволюционного процесса, обуславливающего видообразование [10]. Разработка и использование CAPS маркеров способствует более точному картированию генов интереса, в том числе устойчивости к патогенам [11, 12].

RAPD (random amplified polymorphic DNA; амплифицируемые ДНК фрагменты) – метод амплификации ДНК сегментов с использованием случайных праймеров (примерно 10 нуклеотидов) не требует предварительного знания последовательности ДНК, однако вследствие стохастической природы ДНК амплификации важна оптимизация и поддержание соответствующих условий для получения воспроизводимых результатов. Метод используется, например, для изучения близкородственных видов и молекулярной идентичности растений, размножаемых *in vitro* [13, 14, 15]. Недостатками метода являются: доминантность, не очень хорошая воспроизводимость и необходимость высокоочищенной неконтаминированной ДНК, поскольку использование коротких случайных праймеров может приводить к

амплификации фрагментов различных организмов; локус-неспецифичность маркеров и негомологичность фрагментов одинакового размера (homoplasy). Вследствие недостаточной воспроизводмиости RAPDs сложно использовать в межлабораторных исследованиях. Вследствие вышеупомянутого значимость получаемых результатов и их интерпретация могут подвергаться сомнению.

На основе полученных с использованием случайных праймеров RAPD бэндов, их разделения, экстрагирования, клонирования и секвенирования и дизайна специфических праймеров разработан тип маркеров ***SCAR*** (sequence characterized amplified region, характерная последовательность амплифицируемого участка), развитие и использование которых экспоненциально растет с начала 1990-х годов. Так, разрабатываются специфические SCAR маркеры, позволяющие дискриминировать определенные молекулярные фенотипы различных видов одного рода [16], используемые в таксономии [17] дискриминирование экотипов [18], выявление уникальных сомаклональных вариантов и молекулярных событий, сопряженных с вариабельностью сомаклонов [19]. Подобный тип маркеров обладает потенциалом, в селекции [20] позволяет различать определенные патогенные штаммы микроорганизмов [21] и др.

AFLP (amplified fragment length polymorphism; амплификация полиморфных по длине фрагментов ДНК) – метод селективной ПЦР амплификации полиморфных по длине фрагментов, полученных в результате энзиматического расщепления геномной ДНК эндонуклеазами рестрикции. К полученным после рестрикции фрагментам для селективной амплификации «пришиваются» олигонуклеотидные адаптеры. Таким образом, праймеры состоят из адаптера и последовательности нескольких нуклеотидов (1-5), специфичных для узнавания ферментом рестрикции. Полиморфизм фрагментов по длине (обычно несколько десятков на реакцию), обусловленный множеством геномных сайтов рестрикции, выявляется при электрофорезе. Воспроизводимость и высокое количество информативных фрагментов в реакции позволяет использовать метод в различных аспектах: изучение генетической идентичности, идентификация сортов и клонов, филогенетические связи, картирование, MAS и др. Реализация метода требует высокого качества ДНК. Хотя использование AFLPs, как и RAPDs, не требует предварительного знания ДНК последовательностей, эти типы маркеров сложно использовать при изучении различных популяций или даже видов. Недостатками AFLPs являются доминантность аллелей, возможная негомологичность одинаковых по длине фрагментов – гомоплазия. В случае вставки между соседними местами рестрикции наблюдается утрата короткого фрагмента и появление более длинного. При использовании доминантных маркеров, например, в диплоидах один бэнд имеет место в случаях, если одна или обе гомологичные хромосомы содержат амплифицируемую последовательность, т.е. невозможно различить гомо- и гетерозиготы. В полиплоидах неясность усугубляется, поскольку даже в случае выявления определенного бэнда тем более трудно сказать, в каком количестве аллель присутствует [22]. Тем не менее, использование AFLP маркеров в течение последнего десятилетия (PubMed) практически не снижалось. Метод используется при изучении эволюционных и экологических аспектов живых организмов, при сохранении видов.

Микросателлиты (***SSR***, simple sequence repeat; тандемы повторов простых последовательностей) – участки ДНК, состоящие из тандемов повторяющихся единиц: моно-, ди-, три-, тетра- или пента-нуклеотидов [23]. Микросателлиты присутствуют как в некодирующих, так и в кодирующих областях генома, а также в хлоропластном [24] и митохондриальном геномах [25]. Естественными причинами разнообразия в количестве повторов единиц микросателлитов в геноме являются «проскальзывание» (slippage) полимеразы в ходе репликации ДНК, и/или несоответствующий кроссинговер, несовпадение/восстановление повреждений двойной нити ДНК, а также перемещения ретротранспозонов. Эти вариации приводят к полиморфизму по длине фрагментов, выявляемых при электрофорезе [26]. Этот тип генетических маркеров приобрел в последнее десятилетие большую значимость благодаря комплексу свойств: гипервариабельность, мультиаллельная природа, кодоминантное наследование, высокая воспроизводимость, относительное обилие, экстенсивное распределение по геному, высокая пропускная способность, податливость автоматизации процесса. Микросателлиты используются при изучении филогеографии, структуры популяций, сортовой идентификации, выявлении источника происхождения (отцовства).

По сравнению с SSR использование RAPD недостаточно воспроизводимо, тогда как RFLP не обладают высокой пропускной способностью, а AFLP осложнен тем, что индивидуальные полосы могут состоять из нескольких фрагментов, идентичных по размеру, особенно в больших геномах. Хотя и для SSR гомоплазия, связанная с высокой скоростью мутаций и возможностью обратных мутаций, может быть определенной проблемой. Одним из недостатков SSR маркеров является стоимость их разработки, поскольку для создания праймеров необходимо клонирование и секвенирование произвольных участков ДНК, выявление микросателлитных повторов и определение фланкирующих областей таких участков генома для определения специфического локуса. Микросателлиты могут быть разработаны на основе геномных библиотек ДНК или библиотек, обогащенных для специфических микросателлитов. Некоторые микросателлиты могут быть не просто ди-, три-, или тетра-повторами, а составными повторами, в которых возможен полиморфизм по одному нуклеотиду. Эти аллели различаются на 1 пару нуклеотидов и требуют особого внимания при генотипировании. Аллельные различия в размерах микросателлитов, на которых и основано их использование, в то же время могут обусловить ошибки при амплификации и, соответственно, ошибочной интерпретации генотипирования при оценке генетического разнообразия, структуры популяций, родословной

и др. Исследователи показали, что количество ошибок при использовании микросателлитов прямо коррелирует с размером ПЦР продукта [27].

EST-SSR (*ESTs,* expressed sequence tag, короткие фрагменты последовательностей клонированных участков ДНК (mRNA→cDNA)), которые используются для идентификации транскриптов генов, могут быть источниками для выявления SSRs, т.е. для поиска микросателлитов в экспрессируемых участках генома [28]. Поскольку EST последовательности консервативны, межвидовая EST-SSRs ПЦР амплификация, возможно, более успешна, чем SSRs геномной ДНК. Этот тип маркеров представляет интерес, поскольку их разработка недорога, они отражают транскрибируемые гены, чьи предполагаемые функции зачастую могут быть определены с помощью выявления гомологии. Но качество SSR-EST последовательностей очень важно, поскольку дизайн праймеров может быть осложнен рядом обстоятельств: распространением одного или обоих праймеров на участки сплайсинга; наличием больших интронов в последовательности геномной ДНК, использованием «неполноценной» (questionable) информации относительно создания праймера, и созданием праймеров для химерных cDNA клонов. Таким образом, дизайн праймеров успешен на 60-90%. К началу 2013 г. в базах данных (GenBank) доступны более 74 млн. ESTs. Однако создание генных микросателлитных маркеров ограничено теми видами, для которых имеется достаточно большое количество ESTs. С помощью определенных компьютерных программ данные последовательностей ESTs, генов и cDNA клонов могут быть загружены из GenBank и сканированы на выявление SSRs [табл. 1 в 28]. Приложения использования данного типа маркеров – функциональный геном, ассоциативное картирование, анализ генетического разнообразия, анализ количественных признаков и др. EST-SSR маркеры могут способствовать эволюционному анализу широкого представительства таксонов, анализу видов, ресурсы которых крайне ограничены [29]. Продуктами EST-SSR маркеров являются четкие бэнды, отчетливые аллельные пики [28]. Поскольку эти маркеры являются производными транскриптов, они могут быть использованы для оценки функционального разнообразия естественных популяций или коллекций гермоплазмы. Особенно в случаях маркирования генов, ответственных за фенотипические признаки, эти маркеры полезны для использования в селекции. Как было упомянуто выше, на основе EST-последовательностей разрабатываются и CAPS маркеры.

ISSR (inter-simple sequence repeat, область генома между двумя соседними, противоположно ориентированными микросателлитами) – в качестве праймеров используется последовательность сердцевинной части микросателлита с несколькими (1-3) нуклеотидами, примыкающими к тандему повторностей. Десятки фрагментов множества локусов, полученных в ПЦР, разделяются электрофорезом и оцениваются на присутствие или отсутствие (вследствие доминантности маркеров) фрагментов определенного размера. Главное преимущество данного типа маркеров – отсутствие необходимости знания последовательностей при конструировании праймеров. Следует учитывать, что вследствие мультилокусности возможна гомоплазия. Маркеры данного типа используются для выявления генетической идентичности, родословной, дифференциации клонов, микроклонов и линий, таксономии близкородственных видов.

SNP (single-nucleotide polymorphism; полиморфизм по одному нуклеотиду) – маркеры данного типа становятся все более используемыми в исследованиях генома. Техника основана на том, что в организмах изменения в одном нуклеотиде приводят к точечным мутациям, обуславливая тем самым полиморфизм по одному нуклеотиду (диаллельный тип маркеров). Для создания специфических праймеров необходимо знание последовательностей и фланкирующих областей. Несмотря на высокую стоимость, использование метода в последние годы возрастает. Поскольку метод позволяет автоматизированно осуществлять высокоразрешающее генотипирование с одновременным использованием большого количества SNP маркеров; присутствие многих тысяч проб на чипе позволяет одновременно анализировать множество SNPs. Но в то же время различие аллелей только по одному нуклеотиду и множество проб делает невозможным создание оптимальных условий гибридизации для всего массива проб. Это в ряде случаев приводит к гибридизации анализируемой ДНК с несоответствующими пробами. Оценка большого объема данных становится серьезной проблемой, так как не представляется возможным предпринять point-by-point оценку обоснованности тех или иных данных [30].

Поскольку применение SSR и SNP маркеров постоянно возрастает, исследователи провели сравнительную характеристику этих двух типов маркеров [31]. Так SSRs имеют следующие преимущества перед SNPs. В SSR локусах превышение на определенное количество повторностей может рассматриваться как полиморфизм, в то время как для идентификации SNPs гомологичных областей должны быть сиквенированы области многих хромосом. SSRs обладают несущественной необъективностью, т.е. оценка полиморфизма с их помощью не зависит от исходного перечня сортов. Успех SSR амплификации близкородственных сортов выше, чем для SNP. Локусы SSRs более действенны для выявления смесей, чем SNPs. Достоверность SSR генотипирования оценить легче, поскольку бóльшая доля ошибок может быть выявлена при анализе родословной, когда имеет место много аллелей на локус, тогда как для биаллельных SNPs многие ошибки при анализе не выявляются, поскольку они следуют правилам Менделевского наследования. Очевидно, имеют место и недостатки SSR по сравнению с SNP. Так, большое количество аллелей на SSR локус подразумевает анализ большого количества образцов. Более частые спонтанные SSR мутации внутри родословной потенциально осложняют реконструкцию происхождения; обратные мутации затрудняют описание длительной истории популяций. Вариабельность высокополиморфных микросателлитов

может некорректно отражать геномное разнообразие. Микросателлиты требуют включения контролей, в то время как использование SNPs может осуществляться параллельно в ряде лабораторий без необходимости калибровки результатов.

Очевидно, что разработка и использование различных типов молекулярных маркеров основаны на выявлении изменений в геномных последовательностях, что в свою очередь предполагает, в том числе, экстинцию (выпадение) и последующую вставку последовательностей ДНК. Причиной этих событий могут быть ретротранспозоны (РТ). РТ являются неотъемлемой и довольно значительной частью генома всех живых организмов. Маркеры на основе РТ успешно используются для анализа генетических взаимосвязей, филогенеза, видового разнообразия, при создании генетических карт и идентификации генов [32, 33, 34, 35]. Вездесущая природа РТ и вовлеченность в создание геномного разнообразия посредством интеграции больших сегментов ДНК в рассредоточенные локусы хромосом делают их идеальными для создания на этой основе нескольких типов молекулярных маркеров. Так, ***IRAP*** (inter-retrotransposon amplified polymorphism) генерирует ПЦР продукты между двумя близлежащими РТ. ***REMAP*** (retrotransposon-microsatellite-amplified polymorphism) использует один LTR (long terminal repeat) праймер и другой «присоединенный» к 3'концу SSR последовательности, выявляя РТ, интегрированные поблизости этой SSR последовательности. ***iPBS*** амплификация основана на присутствии в сайте связывания праймера обратной транскриптазы в области LTR РТ. ***SSAP*** (sequence-specific amplified polymorphism) является производным AFLP, амплифицируя продукты между местом интеграции РТ и местом рестрикции, к которому «привязан» адаптер.

ТИПЫ, ПРИЧИНЫ И СЛЕДСТВИЯ ОШИБОК ГЕНОТИПИРОВАНИЯ

Очевидно, что достоверность заключений по любому эксперименту по генотипированию определяется качеством полученных результатов. Ошибки генотипирования могут быть обусловлены различными причинами. Их класификация приведена в обзорах [27, 36, 37, 38]. Минимизирование ошибок возможно при оценке особенностей используемых экспериментальных методов, их оптимизации, надлежащему использованию контролей, повторностей, отбору маркеров, а также разработки статистических подходов для выявления ошибок. Так, повышение вероятности ошибок связано с маркерами с более высокой гетерозиготностью, с большим количеством аллелей, большим количеством «stutter» бэндов и большими размерами продуктов [27]. Наличие статтерных бэндов затрудняет оценку гетерозиготности, например, когда две аллели гетерозиготы отличаются только на один повтор, статтеры их перекрывают, образуя смежные бэнды подобной интенсивности.

Четыре типа ошибок генотипирования создают тенденцию к увеличению доли гомозиготности: выпадение аллелей, нуль аллели, неправильная оценка близлежащих аллелей как «stutter» бэндов, доминирование коротких аллелей, когда бóльшие по размеру аллели имеют слабый сигнал ниже порога обнаружения [38]. Ошибки генотипирования могут проявляться в нарушении Менделевского наследования, но есть такие, при наличии которых результат находится все же в соответствии с Менделевским наследованием. Хотя выявить последние гораздо сложнее, они могут оказывать серьезное влияние на достоверность статистического анализа. Даже выявленные подобные ошибки часто могут быть отнесены к более, чем одному источнику внутри родословной [39]. Так SNPs являются биаллельными маркерами, и их наследование соответствует Менделевскому [40], тем не менее, данный тип маркеров может быть причиной большей доли ошибок. Более того, все большие массивы данных содержат ряд ошибок, обусловленных оплошностями исследователя, недостатками программного обеспечения или просто биохимическими аномалиями, что еще усугубляет использование таких высокопроизводительных методов как SNP [41].

Ошибки генотипирования могут приводить к некорректности генетических карт при изучении картирования признаков интереса, некорректности определения родословной, очевидно их влияние на генетический анализ популяций. Важность выявления ошибок генотипирования становится все более очевидной с развитием международных программ, например, по изучению биоразнообразия и необходимостью обработки огромных массивов данных, получаемых в ряде лабораторий, когда достоверность и воспроизводимость результатов приобретают определяющее значение.

Ошибки разделены исследователями на категории [37].

1 - обусловленные особенностями собственно последовательности молекулы ДНК. В случае мутаций, локализованных в комплементарной последовательности одного из маркеров, амплификация нарушается; вставка или делеция близко к микросателлиту может иметь следствием одинаковый размер фрагментов или «считывание» только одного аллеля вместо двух.

2 - низкое качество и/или недостаточное количество ДНК. Как низкое качество, так и малое количество ДНК обуславливают «выпадение» (dropout) аллелей в гетерозиготах (амплификация только одной предпочтительно более короткой аллели). При контаминации образцов ДНК возможна амплификация контаминантных аллелей. Одним из решений проблемы в случае малого количества ДНК предложен метод разделения образца на несколько пробирок и проведение реакции амплификации в каждой из них [42].

3 - артефакты, обусловленные реагентами. Ошибки, связанные с особенностью Tag-полимеразы добавлять к 3´-концу не принадлежащий к последовательности амплифицируемого фрагмента нуклеотид

(как правило, аденин), в результате может появиться дополнительный бэнд (пик). Эта реакция чувствительна к последовательности 5´-конца праймера и длительному времени элонгации в ПЦР. «Slippage» («запинание») Tag-полимеразы на первых стадиях ПЦР приводит к появлению ложных аллелей. Реагенты низкого качества могут отрицательно влиять на условия ПЦР, неадекватные условия электрофореза могут приводить к искаженной картине распределения аллелей по размеру и т.д.

4 - человеческий фактор. Существенная доля ошибок генотипирования обусловлена именно этим фактором [36]. Так, это может быть связано с субъективностью оценки электрофореграмм и радиоавтограмм, и соответствующим некорректным обозначением аллелей, что в свою очередь зависит и от качества данных; загрязнение экзогенной ДНК или случайное перекрестное «заражение» образцов, использование неоптимальных протоколов, неоптимальных праймеров, температуры плавления в ПЦР, ошибки при вводе и обработке данных и др. [43]. На практике для выявления ошибок генотипирования проводится повторная реакция амплификации и сравнение полученных продуктов первоначального и повторного анализа для определенного ряда образцов, другими словами – анализ дублированных образцов.

Количество (коэффициент, процент) ошибок генотипирования оценивается в различных экспериментах на аллель, на локус, на ПЦР, на мультилокус. Но, поскольку ошибки неслучайно распределены по аллелям, локусам, ПЦР, простое сравнение процентов ошибок между ними некорректно [38]. Так, при сравнении процента ошибки на локус выпадение аллели в гомозиготном локусе не проявляется, а выпадение аллели в гетерозиготном будет выглядеть как гомозигота. Поэтому аллельное выпадение в гомозиготе и гетеролокус с утраченной аллелью будут оценены одинаково и, следовательно, процент ошибки между локусами или популяциями, которые реально различаются в степени гетерозиготности, оценивается некорректно. Наиболее приемлемым считается определение доли ошибки на локус. Процент ошибки для определенной аллели или локуса дает определенную информацию на их целесообразность «исключения» из эксперимента для повышения достоверности результатов. Какой процент ошибки может быть значим для корректности оценки результатов – видно из примера: при 1% ошибки в обозначении аллели только в 62% случаев повторное сравнение генотипа индивидуума подтверждало идентичность, при 2% показатель снижался до 40% [27]. В моделируемых исследованиях было продемонстрировано, что ошибка в 3% отрицательно сказывалась на выявлении неравновесности сцепления признаков (linkage disequilibrium, LD), что мешает, например, выявлению взаимосвязи в комплексах генов, ответственных за болезнь [44].

Согласно определению исследователей [36], следует помнить, что потенциальные последствия ошибок на выводы обратно пропорциональны масштабам выводов. Универсального подхода для исключения ошибок генотипирования не существует, хотя бы по целому ряду объективных причин их природы. Тем не менее, исследователи [37] формулируют определяющие установки для уменьшения количества ошибок и снижения их влияния на конечный результат. В их числе – оценка качества предполагаемых к анализу образцов (качество ДНК) и уровень технических возможностей для реализации; проведение пилотных экспериментов для сравнительной оценки теоретического и реального коэффициента ошибки; контроль качества должен осуществляться на всех этапах исследования для выявления возможного больших вариантов ошибок. Экономия материальных и трудовых затрат должна быть подчинена снижению количества ошибок и повышению эффективности анализа полученных результатов. Только высококвалифицированный персонал, руководствуясь и используя стандарты качества проведения процедур тестирования, с возможно меньшими произвольными манипуляциями, с возможно большей автоматизацией процессов, способен минимизировать процент ошибок генотипирования.

Разработан ряд подходов для выявления ошибок генотипирования и определения их процента, которые включают сравнение дублированных образцов, независимое обозначение размеров аллелей, проверку соответствия Менделевскому наследованию [45]. Но ни один из подходов не способен выявить все ошибки. Исследователи сравнили коммерческий набор микросателлитов и специально созданный высокоразрешающий набор для картирования (commercial genomewide set and custom-designed fine-resolution mapping set). Создание commercial genomewide set основано как на локализации, так и на способности амплифицировать ДНК в обычных условиях. Поэтому все проблемные праймеры заменяются на оптимальные. При конструировании праймеров для fine-resolution mapping set маркеры отбираются согласно их хромосомной локализации, и их амплифицирующая способность не всегда оптимальна. Эта особенность отразилась в первоначальных неудачах ПЦР. Так, ошибки Менделевского наследования составили 0,13% для первого набора праймеров и 1,19% для второго из-за трудностей подсчета некоторых аллелей во втором случае. Проверка исходных образцов и дубликатов (concordance checking) выявила только ошибки, обусловленные человеческим фактором: пропущенные аллели, обозначенные ошибочно и перепутанные образцы, тогда как Менделевское наследование показало дополнительные ошибки, такие как мутации и нуль аллели. Последний подход был более достоверен при использовании обоих типов наборов микросателлитов, тогда как специальный набор маркеров для картирования был более успешен при сравнении дубликатов образцов.

Самый очевидный из подходов – повторное генотипирование и сравнение репрезентативного ряда образцов, хотя и связан с достаточно большими дополнительными экспериментальными затратами, все же не дает полной гарантии достоверного результата. Наиболее часто используемый статистический тест при

анализе популяций – определение соответствия/отклонения равновесию Харди-Вайнберга (HWE) (Deviations from Hardy–Weinberg equilibrium, HWD) [46]. Принцип Харди-Вайнберга описывает, каким образом вариация в популяции поддерживается на основе Менделевского наследования. Закон Харди-Вайнберга гласит, что частота аллелей (т.е. вариаций генов) и генотипа в популяции, благодаря Менделевскому наследованию, в отсутствие эволюционных воздействий остается константной в поколениях. Ряд допущений необходимы для соблюдения этого принципа: диплоидные организмы, только половое размножение, поколения не перекрываются, не дублируются, случайные скрещивания (при инбридинге увеличивается гомозиготность всех генов), неопределенно большой объем популяции, нет миграций, мутаций и отбора.

В локусе с двумя аллелями $(p+q)=(p+q)^2=p^2+2pq+q^2=1$, т.е. пропорции генотипов, согласно Харди-Вайнбергу, равны $p^2+2pq+q^2$ (биномиальное распределение).

Для более чем двух аллелей $(p+q+r)^2=p^2+q^2+r^2+2pq+2pr+2qr$.

В более общем виде для аллелей $A_1, \ldots A_n$ с частотой аллелей от p_1 до p_n $(p_{1+\ldots+}p_n)^2$.

Отклонения от равновесия Харди-Вайнберга могут свидетельствовать об ошибках генотипирования. Но при этом следует помнить, что отклонения могут быть обусловлены и самой структурой популяции. Что, собственно, и важно выявить и разграничить.

Для оценки того, насколько анализируемые данные совпадают с теоретически ожидаемыми, используется «критерий согласия» Пирсона, метод *хи-квадрат* ($\chi 2$). Сравниваются две совокупности: одна – эмпирическое распределение частот, другая представляет собой выборку с теми же параметрами (*n, M, S* и др.), что и эмпирическая, но ее частотное распределение построено в точном соответствии с выбранным теоретическим законом, которому предположительно подчиняется поведение изучаемой случайной величины. В общем виде формула критерия соответствия может быть записана следующим образом:

$$\chi 2 = \sum \frac{(\alpha - A)2}{A},$$

где α – фактическая частота наблюдений;

A – теоретическая ожидаемая частота для данного класса.

К сожалению, проверка на несоответствие Менделевскому наследованию не исключает всех ошибок генотипирования [40]. Исследователями с использованием различных подходов, в том числе TaqMan, RFLP, секвенирование, масс-спектроскопия, SNPs (распределенных как в генных, так и в межгенных областях), было проанализировано 107000 генерированных генотипов. Из 313 SNPs, частоты минорных аллелей которых были в пределах 0,06–0,49, в 36 случаях (11,5%) наблюдались достоверные отклонения от HWE. Для 21 SNPs были выявлены ошибки генотипирования; для 5-ти SNPs – неспецифическое связывание; для 10-ти SNPs причины отклонения не были выяснены. Согласно каждой из использованных технологий и уровню значимости отклонений от HWE ($0,01<P<0,05$ или $P<0,01$), был сделан вывод, что источниками ошибок, по крайней мере отчасти, явилась используемая SNP методология. Результат ретроспективного анализа этих экспериментов по идентификации доли неспецифических проб и генотипических ошибок позволил авторам развить и стандартизировать процесс и в дальнейших исследованиях по генотипированию 96-ти образцов с использованием 1434 SNPs, снизить HWE до 10 ($P<0,01$).

Можно полагать, что незначительный процент ошибок при генотипировании популяций с использованием микросателлитов при значительном количестве образцов и большом количестве аллелей не окажет серьезного влияния на характеристику структуры популяции. Однако было установлено, что ошибки, обусловленные особенностями образцов, могут существенно повлиять на результат, несмотря на большую выборку [38]. Для оценки влияния отдельных образцов на равновесие Харди-Вайнберга использовался «a jackknife» анализ, который осуществлялся с использованием соответствующих программ. Каждый образец последовательно удалялся из общего массива данных и для оставшихся образцов рассчитывался показатель Харди-Вайнберга для всех аллелей. «Влиятельными» образцами считались те, при удалении которых уровень значимости становился $P>0,05$, тогда как для всей суммы образцов был $P\leq 0,05$. Было выявлено 33 случая, когда удаление образца изменяло статус локуса, как влияющего на отклонение равновесия на находящийся в равновесии ($P>0,05$).

Величина влияния определялась преобразованием величины *P:*

$$odds = \frac{p}{1-p}.$$

Таким образом, было показано, что равновесие Харди-Вайнберга было очень чувствительно к гомозиготности редких аллелей в отдельных индивидуумах, и более 50% этих случаев было обусловлено ошибками генотипирования образцов низкого качества. По определению исследователей это указывает на возможность того, что даже «нормальный» уровень лабораторных ошибок может обусловить переоценку влияния отдельных маркеров на равновесие Харди-Вайнберга и тем самым на оценку структуры популяции, в том числе «завышения» роли инбридинга. Любая ошибка, которая существенно влияет на частоту

распределения, способствует некорректному отражению степени дифференциации популяции. Например, если аллель бóльшего размера имеет место только в одной из двух популяций, ошибка, вызывающая ее выпадение (dropout), делает эти две формации более схожими. Наоборот, если одна и та же аллель присутствует в двух популяциях, ее выпадение, снижая ее частоту, приводит к тому, что значительность дифференциации определяется вторичными различиями в частотах редких аллелей.

Еще в начале 90-х годов при изучении генома человека появились данные о неамплификации отдельных аллелей, например, 7 из 23 аллелей (30%) на хромосоме 16 не амплифицировались у некоторых представителей родословной [47]. При генотипировании *Cervus elaphus* не амплифицировались 3 аллели из 16, что было выявлено по несовпадению в парах мать-потомство и существенным отклонением от равновесия Харди-Вайнберга [48]. Анализ двух из трех аллелей показал, что исключение даже нескольких неамплифицируемых гомозигот может иметь серьезный эффект на интерпретацию распределения частот в генотипе и неправильной оценке инбридинга в популяции. Поскольку многие естественные популяции проявляют Харди-Вайнберг неравновесие, как следствие инбридинга или смешивания, необходимо дифференцировать эти случаи и отклонения от равновесия, обусловленные нуль аллелями. Исследователями был предложен алгоритм расчета, который может быть использован для разделения локусов, проявляющих нуль аллели от таковых, отражающих биологический сигнал инбридинга [49].

В настоящее время явление неамплификации микросателлитных аллелей уже признано однозначно. Нуль аллели выявляются не только при использовании микросателлитов, но также и EST-SSR [28]. Молекулярная природа нуль аллелей обусловлена мутациями: заменами, вставками, делециями. Связь между наличием нуль аллелей и высокой вариабельностью фланкирующих областей была зафиксирована в ряде молекулярно-генетических исследований. Так компьютерная симуляция, осуществленная исследователями, и ее приложение к эмпирическим данным позволили заключить, что нуль аллели наблюдались в популяциях с высоким уровнем разнообразия во фланкирующих областях (≥0,001) [50]. Нуль аллели с большой вероятностью фиксировались в популяциях большого эффективного размера с высокой долей мутаций во фланкирующих областях и этим отличались от популяции («центральной»), на основе которой были клонированы аллели и сконструированы праймеры, Другими словами, высокая частота нуль аллелей в исследуемых популяциях была обусловлена высоким уровнем дифференциации между ними и «центральной» популяцией.

Ошибки вследствие «stuttering» (обусловленные особенностями Taq-полимеразы), выпадение бóльших аллелей и нуль аллели, в отличие от стохастических (произвольных, случайных) ошибок, создают последовательные (consistent) отклонения от действительно имеющей место картины аллелей и соответственно отклонения от адекватности генотипирования [51]. Так, размеры и форма «stuttering» образцов варьируют среди локусов, некоторые маркеры проявляют низкий уровень «stuttering», тогда как другие могут образовывать два и более пиков. Интерпретация подобных образцов затруднена, например, соседних аллелей в гетерозиготе, различающихся на один динуклеотидный повтор. Такой аллель гетерозиготы может быть оценен как гомозигота с аллелью бóльшего размера. В целом это приведет к сдвигу в сторону аллелей бóльшего размера, снижая гетерозиготность и увеличивая кажущийся уровень инбридинга соответствующих локусов. Что касается выпадения аллели бóльшего размера, вследствие преимущественной амплификации более короткого аллеля, пик последнего может быть гораздо выше первого, а в образцах низкого качества бóльший аллель не амплифицируется вообще. Выпадение аллелей увеличивается с их размером, а в некоторых локусах даже увеличение концентрации ДНК не снижает уровень выпадения, указывая на технические ограничения при амплификации больших аллелей. Нуль аллели – третья причина последовательных ошибок генотипирования, не выявляются по определению, причина их главным образом в мутации в месте прайминга (присоединения праймера). В случае гетерозиготы образцы будут оценены как гомозигота, а в случае гомозиготы – как отсутствие реакции амплификации вообще. В общем, это приводит к снижению частоты гетерозиготности и повышению кажущегося уровня инбридинга. Исследователи рекомендуют стандартные процедуры и протокол, позволяющие предупредить, снизить уровень ошибок генотипирования при изучении диких популяций [52]. Они включают скрининг (отбор) микросателлитных локусов, повторный анализ части (subset) образцов, оценку полного массива данных, тестирование ошибок, уменьшение ошибок в последующих анализах и предоставление отчетности по частоте ошибок.

В работе [53] при анализе 233 публикаций с упоминаемыми нуль аллелями было обращено внимание на несколько положений: как эти аллели были определены; был ли проведен редизайн праймеров для этих локусов; каким методом определялась их частота; проводилось ли их сиквенирование для выяснения молекулярной основы их «проявления»; были ли локусы с нуль аллелями оставлены или исключены из последующего анализа. Наиболее часто используемый в этих работах подход для выявления нуль аллелей был основан на наблюдаемом дефиците гетерозигот в популяциях. Использовалась также оценка отклонений от Харди-Вайнберг равновесия. Хотя на этот показатель оказывает влияние и структура субпопуляций, инбридинг или селекция, затронувшая место, близкорасположенное к микросателлиту. В нескольких работах был привлечен анализ родословной, что является более весомым свидетельством. Авторы пришли к выводу, что частоты нуль аллелей могут приводить к ложным выводам относительно

родословной. Обусловленная нуль аллелями и наблюдаемая высокая гомозиготность может приводить к завышенной оценке дифференциации популяций. При высокой дифференциации популяций наличие нуль аллелей приводит к завышению индекса фиксации (F_{ST}) – меры дифференциации популяции и генетических расстояний.

Поскольку праймеры, фланкирующие EST-SSRs, являются производными достаточно консервативных последовательностей, вероятно, для этих маркеров нуль аллели могут быть меньшей проблемой. Сравнение 25 EST-SSRs и 17 SSRs маркеров при генотипировании ряда представителей хвойных (*Picea spp.)* показало меньшую гетерозиготность для EST-SSRs маркеров (среднее H=0,65 vs 0,72). Но при этом значения вероятности инбридинга (*F*) для специфических SSR маркеров позволили идентифицировать локусы с мнимыми нуль аллелями, в результате указывая на существенно более низкий очевидный инбридинг (среднее F=0,046 vs 0,126) [54].

Как упоминалось выше, для молекулярных маркеров возможно проявление гомоплазии – негомологичности амплифицированных последовательностей одинакового размера. Использование микросателлитов и секвенирования фрагментов ДНК существенно повысило возможности исследователей по выявлению признаков, имеющих гомопластический характер и их влияния на популяционный анализ [55]. Первоначально концепцию гомоплазии использовали эволюционисты, описывая тот факт, что характерный признак, присутствующий в двух видах, может не иметь одного общего предка, а быть результатом конвергенции, параллелизма или риверсии. Таким образом, гомоплазия имеет значение при оценке филогенетических взаимоотношений между видами. Для молекулярного маркера гомоплазия имеет место тогда, когда различные копии локуса одинаковы по размеру и не разделяются электрофоретически (электроморфы), но не идентичны по происхождению. Например, микросателлиты мутируют с достаточно высокой скоростью и аллели одинакового размера имеют скрытые различия в последовательностях, которые могут быть выявлены только секвенированием или (SSCP, single-strand conformation polymorphism) [56]. Гомоплазия по размеру внутри и между популяциями побуждает исследователей определять ее индекс как вероятность того, что в данном локусе две копии гена одного состояния не идентичны по своему происхождению.

Еще более значимо для интерпретации результатов проявление гомоплазии при AFLP анализе. Так, поскольку доля одинакового размера AFLP профилей составляла от 0 до 5% для близких сородичей и многократно возрастала для удаленных таксонов, исследователи высказали предположение, что изучение филогении с использованием AFLPs возможно только для близкородственных таксонов [57]. Используя AFLP при сравнительной характеристике *Phaseolus lunatus* и *Lolium perenne,* исследователи показали, что для обоих видов распределение по размеру фрагментов асимметрично с гораздо большей долей коротких фрагментов по сравнению с длинными [58]. Меньшего размера фрагменты более подвержены комиграции при электрофорезе и проявлению гомопластичности. Исследователи высказали гипотезу, что уровень гомоплазии по размеру является функцией размера и это может смещать оценки генетического разнообразия, основанного на частотах AFLP фрагментов. Проведя анализ in silico полных геномных последовательностей 14 видов, включая эукариоты, прокариоты и архея, исследователи выявили следующее: показанная ранее положительная корреляция между процентом гомоплазии и размером генома есть прямое следствие количества наблюдаемых бэндов и содержания ГЦ оснований в геноме [59]. Регрессия пропорции гомоплазии на количество бэндов была идентична для всех анализированных видов, несмотря на 1,000-кратную разницу в размерах генома.

Поскольку исследователи признают, что генотипирование с помощью молекулярных маркеров весьма субъективная процедура в зависимости от подготовленности персонала и возможностей конкретной лаборатории, в качестве одного из критериев, например, для AFLP маркеров разработан алгоритм scanAFLP отбора маркеров на основе интенсивности распределения фрагментов среди образцов [60]. Метод сравнен с предложенным ранее AFLPscore [56], в котором в качестве критерия использовались пороги интенсивности флюоресценции продукта AFLP-полимеразной реакции. Алгоритмы использовались для сканирования популяции из 619 индивидуумов *Arabis alpina* из 191 места сбора и оценки генетической структуры, разнообразия и дифференциации. В каждой плашке на 96 ячеек использовали следующие контроли генотипирования, начиная с выделения ДНК: один отрицательный контроль (без образца), два образца включали во все плашки (повторности, положительный контроль для всех плашек) и два образца на каждую плашку анализировали дважды (положительный контроль на плашку). Кроме того, один образец в качестве дубликата собирали в каждом месте сбора. Из них 39 случайно отобранных включали в анализ в качестве «blind» контролей для расчета несовпадения уровня ошибок. Автоматизированная процедура отбора маркеров существенно уменьшила процент ошибок, понизив случайные сигналы, связанные как с индивидуумами, так и популяциями. Случайно отобранные наборы маркеров показали заметно меньшую вариабельность между дубликатами по сравнению с наборами, отобранными согласно определенному критерию. Следует обратить внимание на набор контролей для генотипирования. Подобный подход необходим при использовании любого типа молекулярных маркеров.

Очевидно, что наряду с необходимостью выявления и уменьшения доли ошибок генотипирования, которые имеют место в любом эксперименте при использовании различных типов маркеров, важно

соблюдать компромисс между выбраковкой локусов, порождающих ошибки, и повышением потенциала оставляемых локусов для усиления генетического сигнала популяции. Компромисс может быть различным в различных исследованиях, но главное, чтобы этот сигнал не был потерян в угоду «приемлемого» уровня ошибок, и исследователи предлагают эмпирический подход для достижения желаемого компромисса [61].

ЗАКЛЮЧЕНИЕ

За последние два десятилетия наблюдается экспоненциальный рост использования молекулярных маркеров в изучении разнообразных аспектов жизнедеятельности растительных организмов. Общее количество работ уже приближается к двум десяткам тысяч. В связи с этим очевидна не только значимость использования различных типов молекулярных маркеров, но и корректность оценки получаемых с их помощью результатов. Опираясь на огромный экспериментальный опыт, в целом ряде обзоров исследователями сформулированы преимущества и недостатки используемых молекулярных маркеров, основные требования при их выборе для того или иного рода исследований. Несмотря на разнообразие маркеров, существует ряд первоочередных общих требований и критериев, позволяющих повысить достоверность получаемых результатов. В их числе выбор адекватного типа маркеров для решения поставленной задачи; достаточно высокая полиморфность маркеров, высокая разрешающая способность, кодоминантность, частота и равномерность распределения по геному, высокая воспроизводимость. Не менее важным и существенным при генетическом анализе является минимизирование ошибок, которые разделены исследователями на категории. Главные из них: соответствующее качество и количество анализируемой ДНК; оценка и отбор оптимальных праймеров для повышения информативности соответствующих локусов; особенности Tag-полимеразы и оптимизация условий ПЦР; надлежащее использование повторностей и контролей, а также адекватность статистических подходов для оценки полученных результатов и выявления ошибок при анализе.

Работа выполнена в рамках Гранта 0048/ГФ по подприоритету: «Исследования в области продовольственной безопасности» Бюджетной программы 055, финансируемой Государственным учреждением «Комитет науки Министерства образования и науки Республики Казахстан».

ЛИТЕРАТУРА

1. *Ganal M.W., Polley A., Graner E.M., Plieske J., Wieseke R., Luerssen H., Durstewitz G. Large SNP arrays for genotyping in crop plants // J Biosci. – 2012. – Vol. 37. – P. 821-288.*
2. *Miedaner T., Korzun V. Marker-assisted selection for disease resistance in wheat and barley breeding // Phytopathology. – 2012. – Vol.102. – P. 560-566.*
3. *Paux E., Sourdille P., Mackay I., Feuillet C. Sequence-based marker development in wheat: advances and applications to breeding // Biotechnol Adv. – 2001. – Vol. 30. – P. 1071-1088.*
4. *Hall D., Tegström C., Ingvarsson P.K. Using association mapping to dissect the genetic basis of complex traits in plants // Brief Funct Genomics. – 2010. – Vol. 9. – P. 157-165.*
5. *Zimmer E.A., Wen J. Using nuclear gene data for plant phylogenetics: progress and prospects // Mol Phylogenet Evol. – 2012. – Vol. 65. – P. 774-785.*
6. *Kumar P., Gupta V.K., Misra A.K., Modi D. R., Pandey B. K. Potential of Molecular Markers in Plant Biotechnology // Plant Omics Journal. – 2009. – Vol. 2. – P. 141-162.*
7. *Woodhead M., Russell J., Squirrell J., Hollingsworth P. M., Mackenzie K., Gibb M., Powell W. Comparative analysis of population genetic structure in Athyrium distentifolium (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions // Molecular Ecology. – 2005. – Vol. 14. – P. 1681-1695.*
8. *Kuczmog A., Galambos A., Horváth S., Mátai A., Kozma P., Szegedi E., Putnoky P. Mapping of crown gall resistance locus Rcg1 in grapevine // Theor Appl Genet. – 2012. – Vol.125. – P. 1565-1574.*
9. *Ujihara T, Taniguchi F, Tanaka J, Hayashi N. Development of Expressed Sequence Tag (EST)-based Cleaved Amplified Polymorphic Sequence (CAPS) markers of tea plant and their application to cultivar identification // J Agric Food Chem. – 2011. – Vol. 59. – P. 1557-1564.*
10. *Segatto A.L.A., Caze A.L.R., Turchetto C., Klahre U., Kuhlemeier C., Bonatto S.L., Freitas L.B. Nuclear and plastid markers reveal the persistence of genetic identity: A newperspective on the evolutionary history of Petunia exserta // Molecular Phylogenetics and Evolution. - 2014. – Vol.70. – P. 504–512.*
11. *Cui Y., Lee M.Y., Huo N., Bragg J., Yan L., Yuan C., Li C., Holditch S.J., Xie J., Luo M.C., Li D., Yu J., Martin J., Schackwitz W., Gu Y.Q., Vogel J.P., Jackson A.O., Liu Z., Garvin D.F. Fine mapping of the Bsr1 barley stripe mosaic virus resistance gene in the model grass Brachypodium distachyon // PLoS One. – 2012. – Vol. 7:e38333.*

12. *Jo K.-R., Arens M., Kim T.-Y., Jongsma M.A., Visser R.G.F., Jacobsen E., Vossen H.J. Mapping of the S. demissum late blight resistance gene R8 to a new locus on chromosome IX // Theor Appl Genet. – 2011. – Vol. 123. – P. 1331–1340.*

13. *Shu Q.Y., Liu G.S., Qi D.M., Chu C.C., Liu .J, Li H.J. An effective method for axillary bud culture and RAPD analysis of cloned plants in tetraploid black locust // Plant Cell Rep. – 2003. – Vol. 22. – P. 1751-1780.*

14. *Sreedhar R.V., Venkatachalam L., Bhagyalakshmi N. Genetic fidelity of long-term micropropagated shoot cultures of vanilla (Vanilla planifolia Andrews) as assessed by molecular markers // Biotechnol. J. – 2007. – Vol. 2. – P. 1007–1013.*

15. *Rai G.K, Singh M., Rai N.P., Bhardwaj D.R., Kumar S. In vitro propagation of spine gourd (Momordica dioica Roxb.) and assessment of genetic fidelity of micropropagated plants using RAPD analysis // Physiol Mol Biol Plants. – 2012. – Vol. 18. – P. 273-280.*

16. *Kim J.K., An G.H., Ahn S.H., Moon Y.H., Cha Y.L., Bark S.T., Choi Y.H., Suh S.J., Seo S.G., Kim S.H., Koo B.C. Development of SCAR marker for simultaneous identification of Miscanthus sacchariflorus, M. sinensis and M. x giganteus // Bioprocess Biosyst Eng. – 2012. – Vol.35. – P. 55-59.*

17. *Pankin A.A., Khavkin E.E. Genome-specific SCAR markers help solve taxonomy issues: a case study with Sinapis arvensis (Brassiceae, Brassicaceae) // Am J Bot. – 2011. – Vol. 98:e54-7.*

18. *Ray T., Roy S.C. Genetic diversity of Amaranthus species from the Indo-Gangetic plains revealed by RAPD analysis leading to the development of ecotype-specific SCAR marker // J Hered. – 2009. – Vol. 100. – P. 338-347.*

19. *Osipova E.S., Lysenko E.A., Troitsky A.V., Dolgikh Y.I., Shamina Z.B., Gostimskii S.A. Analysis of SCAR marker nucleotide sequences in maize (Zea mays L.) somaclones // Plant Sci. – 2011. – Vol. 180. – P. 313-322.*

20. *Rahman M., Sun Z., McVetty P.B., Li G. High throughput genome-specific and gene-specific molecular markers for erucic acid genes in Brassica napus (L.) for marker-assisted selection in plant breeding // Theor Appl Genet. – 2008. – Vol. 117. – P. 895-904.*

21. *Naeimi S., Kocsubé S., Antal Z., Okhovvat S.M., Javan-Nikkhah M., Vágvölgyi C., Kredics L. Strain-specific SCAR markers for the detection of Trichoderma harzianum AS12-2, a biological control agent against Rhizoctonia solani, the causal agent of rice sheath blight // Acta Biol Hung. – 2011, Mar. - №62(1). – P. 73-84.*

22. *Falush D., Stephens M., Pritchard J.K. Inference of population structure using multilocus genotype data: dominant markers and null alleles // Molecular Ecology Notes. – 2007. – Vol. 7. – P. 574–578.*

23. *Powel l W., Machray G.C., Provan J. Polymorphism revealed by simple sequence repeats // Trends Plant Sci. – 1996. – Vol. 1. – P. 215-222.*

24. Chung A.M., Staub J.E., Chen J.F. Molecular phylogeny of Cucumis species as revealed by consensus chloroplast SSR marker length and sequence variation // Genome. – 2006. – Vol. 49.

24. *Rajendrakumar P., Biswal A.K., Balachandran S.M., Srinivasarao K., Sundaram R.M. Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions // Bioinformatics. – 2007. – Vol. 23. – P.1-4.*

25. *Kalia R.K. Rai M.K, Kalia S., Singh R., Dhawan A.K. Microsatellite markers: an overview of the recent progress in plants // Euphytica. – 2011. – Vol.177. – P.309–334*

26. *Hoffman J.I., and Amos W. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion // Molecular Ecology. – 2005. – Vol. 14. – P.599-612.*

27. *Varshney R.K., Graner A., Sorrells M.E. Genic microsatellite markers in plants: features and applications // TRENDS in Biotechnology. – 2005. – Vol. 23. – P.48-55.*

28. *Ellis R., JM Burke. EST-SSRs as a resource for population genetic analyses // Heredity. – 2007. – Vol. 99. – P.125-132.*

29. *Saunders I.W., Brohede J., Hannan G.N. Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference // Genomics. – 2007. –Vol. 90. – P. 291-296.*

30. *Guichoux E., Lagache L., Wagner S., Chaumeil P., Le. Ger P., Lepais O., Lepoittevin C., Malausa E., Revardel E., Salin F., Petit R.J. Current trends in microsatellite genotyping // Molecular Ecology Resources. – 2011. – Vol. 11. – P. 591–611.*

31. *Schulman A.H., Flavell A.J., Ellis T.H. The application of LTR retrotransposons as molecular markers in plants // Methods in Molecular Biology. – 2004. – Vol. 260. – P.145–173.*

32. *Kalendar R., Antonius K., Smykal P., Schulman A.H. iPBS: a universal method for DNA Wngerprinting and retrotransposon isolation // Theor Appl Genet. – 2010. –121. – P.1419–1430.*

33. *Castro I., D'Onofrio C., Martin J.P., Ortiz J.M., De Lorenzis G., Ferreira V.O. Pinto-Carnide. Effectiveness of AFLPs and Retrotransposon-Based Markers for the Identification of Portuguese Grapevine Cultivars and Clones Mol Biotechnol. – 2012. - Vol. 52. – P. 26–39.*

34. *Nakatsuka T., Yamada E., Saito M., Hikage T., Ushiku Y., Nishihara M. Construction of the first genetic linkage map of Japanese gentian (Gentianaceae) // BMC Genomics. – 2012. –Vol. 13. – P. 672.*

35. *Bonin A., Bellemain E., Bronken Eidesen P., Pompanon F., Brochmann C., Taberlet P. How to track and assess genotyping errors in population genetics studies // Mol Ecol. – 2004. – Vol. 13. – P.3261-3273.*

36. *Pompanon F., Bonin A., Bellemain E.,Pierre Taberlet. Genotyping Errors: Causes, Consequences And Solutions. Genetics. – 2005. – Vol. 6. – P.847-859.*

37. *Morin P.A., Leduc R. G., Archer F. I., Martien K. K., Huebinger R., Bickham J.W., Taylor B.L. Significant deviations from Hardy–Weinberg equilibrium caused by low levels of microsatellite genotyping errors // Molecular Ecology Resources. – 2009. – Vol. 9. – P.498-504.*

38. *Gordon D., Heath S.C., Ott J. True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms // Hum Hered. – 1999. – Vol.49. – P.65–70.*

39. *Gordon D., Finch S.J., Nothnagel M., Ott J. Power and sample size calculations for case–control genetic association tests when errors are present: application to single nucleotide polymorphisms // Human Heredity. – 2002. – Vol. 54. – P.22–33.*

40. *Sobel E., Papp J. C., Lange K. Detection and integration of genotyping errors in statistical genetics Am. J. Hum. Genet. – 2002. – Vol. 270. – P.496–508.*

41. *Navidi W., Arnheim N., Waterman M.S. A multiple-Tubes Approach for Accurate genotyping of very Small DNA samples by using PCR: Statistical considerations // Am. J. Hum. Genet. – 1992. – Vol. 50. – P.347-359.*

42. *Ghosh, S. et al. Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers // Genome Res. – 1997. – Vol. 7. – P.165–178.*

43. *Hosking L., Lumsden S., Lewis K., Yeo A., McCarthy L., Bansal A., Riley J., Purvis I., a Xu Ch-F. Detection of genotyping errors by Hardy–Weinberg equilibrium testing // European Journal of Human Genetics. – 2004. – Vol. 12. – P.395–399.*

44. *Ewen K.R., Bahlo M., Treloar S.A., Levinson D.F., Mowry B., Barlow J.W., S.J. Foote. Identification and Analysis of Error Types in High-Throughput Genotyping // Am. J. Hum. Genet. – 2000. – Vol. 67. – P.727–736.*

45. *Gomes I., Collins A., Lonjou C. et al. Hardy–Weinberg quality control. Annals of Human Genetics. – 1999. – Vol. 63. – P.535–538.*

46. *Callen D.F., Thompson A.D., Shen Y., Phillips H., Richards Rl., Mulley J.C., Sutherland G.R. Incidence and origin of 'null' alleles in the (AC)n microsatellite markers // Americun Jounrnl of Human genetics. – 1993. – Vol. 52. – P.922-927.*

47. *Pemberton J.M., Slate J., Bancroftt 'D.R, Barrett J.A. Nonamplifying alleles at microsatellite a caution for parentage and population loci: studies // Molecular Ecology. – 1995. – Vol. 4. – P.249-252.*

48. *Van Oosterhout C., Weetman D., Hutchinson W. F. Estimation and adjustment of microsatellite null alleles in nonequilibrium populations // Molecular Ecology Notes. – 2006. – Vol. 6. – P.255–256.*

49. *Chapuis M.-P., Estoup A. Microsatellite Null Alleles and Estimation of Population Differentiation // Mol. Biol. Evol. – 2007. – Vol. 24. – P.621–631.*

50. *Dewoody J., Nason J.D., Hipkins V.D. Mitigating scoring errors in microsatellite data from wild populations // Molecular Ecology Notes. – 2006. – Vol. 6. – P.951–957.*

51. *Dakin E.E., Avise J.C. Microsatellite null alleles in parentage analysis // Heredity. –2004. – Vol. 93. – P.504-509.*

52. *Rungi D., Berube Y., Zhang J., Ralph S., Ritland C.E., Ellis B E., Douglas C., Bohlmann J., Ritland K. Robust simple sequence repeat markers for spruce (Picea spp.) from expressed sequence tags // Theor Appl Genet. – 2004. – Vol. 109. – P.1283–1294.*

53. *Wake D.B., Wake M.H., Specht C.D. Homoplasy: From Detecting Pattern to Determining Process and Mechanism of Evolution // Science. – 2011. – Vol. 331. – P.1032-1035.*

54. *Estoup A., Jarne P., Cornuet J.-M. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis // Molecular Ecology. – 2002. – Vol. 11. – P.1591–1604.*

55. *O'hanlon P.C., Peakall R. A Simple method for the detection of size homoplasy among amplified fragment length polymorphism fragments // Molecular Ecology. – 2000. – Vol.9. – P.815–816.*

56. *Vekemans X., Beauwens T., Lemaire M., Roldcn-Ruiz I. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size // Molecular Ecology. – 2002. – Vol. 11. – P.139-151.*

57. *Caballero A., Quesada H. Homoplasy and Distribution of AFLP Fragments: An Analysis In Silico of the Genome of Different Species // Mol. Biol. Evol. – 2010. – Vol. 27. – P.1139–1151.*

58. *Herrmann D., Poncet B.N., Manel S., Rioux D., Gielly .L, Taberlet P. Gugerli F. Selection criteria for scoring amplified fragment length polymorphisms (AFLPs) positively affect the reliability of population genetic parameter estimates // Genome. – 2010. – Vol. 53. – P.302–310.*

59. *Whitlock R., Hipperson H., Mannarelli M.,. Butlin R .K, Burke T. An objective, rapid and reproducible method for scoring AFLP peak-height data that minimizes genotyping error // Molecular Ecology Resources. – 2008. – Vol. 8. – P.725–735.*

60. *Zhang H., Hare M.P. Identifying and reducing AFLP genotyping error: an example of tradeoffs when comparing population structure in broadcast spawning versus brooding oysters // Heredity. –2012. – Vol. 108. – P. 616–625.*

Биотехнология. Теория и практика. 2013, №4, стр. 20-28
DOI: 10.11134/btp.4.2013.3

ТҮЙІН

ДНҚ полиморфизмін анықтайтын әр түрлі молекулалық маркерлер, қазіргі таңда биологияның әр түрлі саласында қолданылып жүр, соның ішінде генетикалық әр түрлілікті зерттеу және сақтауда, индивидуумдарды идентификациялауда, филогенетикада, стресс факторларға төзімді пайдалы белгілерді карталауда, селекция процесстерінде, биотехнологияда және т.б. қолданылуда. Генотипирлеу кезіндегі алынатын нәтижелерді орынды талдау үшін мынадай фактілерді ескеру қажет, кез келген молекулалық маркерлерді қолдану барысында генотипирлеу қателіктері болуы мүмкін, олардың ең негізгісі - үлкен аллелдердің түсіп қалуы, нөл аллелдер, «stutter», Таг-полимеразалардың ерекшеліктеріне байланысты аллелдер, бірдей мөлшерлі амплифицирленген қатарлардың гомология еместігі (гомоплазия). Зерттеушілер қателіктердің мөлшерлерінің азаюы және соңғы анализге әсерін төмендету үшін айқындалатын жағдайларды тұжырымдайды. Олардың ішінде талданатын ДНҚ-ның сапасы мен мөлшері, қызметкердің техникалық жағдайы және профессионалдылығы, себебі генотипирлеу барысындағы орынсыз нәтижелердің пайда болуының ең негізгі себептерінің бірі адамзат факторының ықпалымен анықталады; теориялық түрде және іс жүзінде қателіктің коэффициентін салыстырмалы түрде бағалау үшін тәжірибе жүргізу. Қателіктерді төмендету маркерлердің скринингісі мен типтерін қолдану ерекшеліктерімен бағаланады; тәжірибелік әдістерді оңтайландыру, бақылауды қолдану арқылы, қайталанулар, сонымен қатар қателіктерді табу үшін статистикалық қадамдарды тексеру арқылы жүзеге асады.

Кілтті сөздер: Молекулалық маркерлер, генотипирлеу, генотипирлеудің қателіктері

SUMMARY

Molecular markers based on the DNA polymorphism have become most significant approach to assess the genetic relatedness/diversity between individuals, population genetics, genetic maps developing, marker assisted selection (MAS), phylogeography, biotechnology and so on. Types of molecular markers and differences in their methodologies are the rationale for the applications of a particular type. However, when using any type of genetic markers are inevitable genotyping errors due to the DNA sequence itself, its quantity and quality, biochemical artifacts and human factors (Pompanon et al., 2005; Hoffman, Amos, 2005; Kumar et al., 2009). Four types of errors leading to erroneous conclusions are the most common ones: allelic dropout, null alleles; misinterpretation of neighboring alleles as stutter; and short allele dominance (Morin et al., 2009). One of the problems, especially the major for AFLP method is size homoplasy. The most obvious of approaches for detecting genotyping errors is the re-genotyping and comparison of a representative number of samples. Statistical tools to take error into account have been developed. The most commonly used statistical test in the analysis of populations is the Hardy–Weinberg equilibrium test on the basis of Mendelian proportions and generally performed using Pearson'schi-squer test, χ2. To reduce the genotyping errors researchers recommend proper use of stringent controls, effective screening of microsatellite loci prior to data collection, reanalysis a subset of samples, combining automated allele calling with programs available to automate scoring with visual inspection of each sample, testing for scoring errors, downstream analyses and reporting error rates (Dewoody et al., 2006).

Keywords: Molecular markers, genotyping, genotyping errors