Математическая биология и биоинформатика
2022. Т. 17. № 2. С. 208–229. doi: 10.17537/2022.17.208

========== **БИОИНФОРМАТИКА** ==========

Мультиплексный *in silico* RAPD-анализ для баркодирования геномов

Кирьянова О.Ю. [*1], Кирьянов И.И.[2], Кулуев Б.Р.[3], Гарафутдинов Р.Р.[3], Чемерис А.В.[3], Губайдуллин И.М.[1,4]

[1]*Уфимский государственный нефтяной технический университет, Уфа, Россия*
[2]*ООО «Корнинг СНГ», Санкт-Петербург, Россия*
[3] *Институт биохимии и генетики – обособленное структурное подразделение Федерального государственного бюджетного научного учреждения Уфимского федерального исследовательского центра, Уфа, Россия*
[4] *Институт нефтехимии и катализа – обособленное структурное подразделение Федерального государственного бюджетного научного учреждения Уфимского федерального исследовательского центра, Уфа, Россия*

Аннотация. В данной работе предлагается новый метод идентификации живых организмов с помощью мультиплексной полимеразной цепной реакции с произвольными праймерами *in silico* (мультиплексный *in silico* RAPD-анализ). Представлены результаты компьютерного моделирования поиска возможных мест отжига праймеров в геномной ДНК с образованием ампликонов определенного диапазона длин (от 51 до 500 нуклеотидов). Полученные данные предложено переводить в бинарный формат, а также в формат геномного штрихкода, в совокупности с используемыми праймерами оказывающийся уникальным для исследуемых геномов. Набор праймеров и полученный штрихкод позволяют однозначно идентифицировать организмы, которым они принадлежат. Проведен сравнительный анализ полученных геномных штрихкодов видов близкородственных растений с целью установления их филогенетических связей, систематизации, классификации, а также ДНК-паспортизации (каталогизации) на уровне отдельных сортов и линий. Предложенный метод позволяет прогнозировать успешность проведения мультиплексной полимеразной цепной реакции с произвольными праймерами в лабораторных условиях. Разработанная программа для поиска мест отжига праймеров является вспомогательным инструментом для предварительного планирования «мокрых» экспериментов по выявлению мультилокусного полиморфизма ДНК и позволяет при определенных допущениях определить «удачные» праймеры для проведения полимеразной цепной реакции, а также определить оптимальное количество праймеров для получения наиболее информативных штрихкодов. Важным моментом является то, что предложенная технология позволяет проводить анализ нуклеотидной последовательности всей геномной ДНК различных эукариотических организмов, а не отдельных фрагментов генома, что делает предложенный метод баркодирования универсальным для всех геномов независимо от их видовой принадлежности.

Ключевые слова: *геномное баркодирование, цифровизация данных, мультиплексная ПЦР, RAPD-анализ, компьютерное моделирование, геном.*

[*]olga.kiryanova27@gmail.com

ВВЕДЕНИЕ

Биологическое разнообразие чрезвычайно велико, но, чтобы его изучать и сохранять, требуется систематизация данных по нему, в том числе на уровне ДНК. При этом негативное влияние человеческой деятельности на популяции живых организмов в ряде мест, ведущее к их исчезновению, повышает актуальность разработки новых методов оценки биоразнообразия [1]. Кроме того, развитие технологий селекции и создание новых сортов растений (а также пород животных, рас грибов, штаммов микроорганизмов) требует более четкой их классификации, информации о происхождении сортов, что может быть установлено с помощью имеющего место полиморфизма ДНК. Существует немало методов выявления полиморфизма ДНК, однако многие из них или слишком громоздки и дорогостоящи, либо не обеспечивают полноты данных, особенно, если они основаны на анализе лишь одного или нескольких отдельных генов. Поэтому важна разработка удобного способа классификации живых организмов на уровне всего генома, получаемые данные с помощью которого могут храниться независимо от самих биологических объектов. Поскольку к настоящему времени полностью секвенированы и депонированы в различные базы данных геномы многих хозяйственно-ценных видов растений и животных, то появляется возможность проводить их геномное баркодирование, служащее затем с одной стороны для классификации этих организмов, а с другой – позволяя прогнозировать удачный исход так называемых «мокрых» экспериментов с конкретным комплектом произвольных праймеров, отбраковывая неудачные, и тем самым экономя время и деньги при анализе отдельных организмов (линий, сортов) какого-либо вида растений, например.

Подход к систематизации и идентификации живых организмов на основе некоторых эволюционно консервативных генов путем присвоения им штрих-кодов был предложен почти два десятилетия назад [2]. Данный метод известен как DNA barcoding или ДНК-баркодирование и служит для молекулярной идентификации, устанавливая принадлежность организма к тому или иному таксону по определенным коротким генетическим фрагментам [3], которыми изначально для животных объектов служили участки митохондриального гена цитохром с-оксидазы [4]. В этом случае проводилась полимеразная цепная реакция (ПЦР) с праймерами, приходящимися на консервативные последовательности и ограничивающими нужный фрагмент ДНК, который затем секвенировался. На основе установленной нуклеотидной последовательности формировался ДНК-штрихкод, лишь отдаленно напоминающий привычный штрихкод и представляющий собой по сути ту же самую последовательность нуклеотидов, отображенную в виде цветных вертикальных штрихов соответствующего конкретным нуклеотидам цвета (A – зеленый, C – синий, G – черный, T – красный). При использовании данного метода есть проблемы с организмами, у которых есть внутриклеточные симбионты, поскольку те имеют свой ген цитохром с-оксидазы с иной последовательностью. Позже выяснилось, что для растений и грибов требуются другие маркерные гены [5, 6]. Более того, для аллополиплоидных видов растений, которыми, в частности, являются возделываемые мягкая и твердая пшеницы, несущие соответственно субгеномы B, A, D и B, A с отличающимися гомологичными генами, требуются отдельные ДНК-штрихкоды, которые с помощью используемой технологии не так просто сформировать. Не всегда для разграничения родов и видов в ДНК-баркодировании работают установленные (в процентах) критерии. К тому же на уровне сортов растений или пород животных данный метод не способен выявлять их генетические отличия, поскольку гены, берущиеся в исследование, высоко консервативны.

Помимо таксономии, в последние годы описано использование метода ДНК-баркодирования при анализе пищевых продуктов на предмет их фальсификации [7]. В одной из работ предложен метод идентификации видов *Anopheles minimus* и *An.*

harrisoni с помощью коротких фрагментов ДНК с целью быстрого выявления возбудителей малярии [8]. Для исследования филогенетического сходства было применено построение специфических штрихкодов, основанное на сравнении коротких нуклеотидных последовательностей *k*-меров длиной от 16 до 22 нуклеотидов [9, 10].

Нами предложен новый вариант баркодирования, названного геномным, чтобы исключить путаницу с описанным выше методом ДНК-баркодирования. Наш метод основан в первую очередь на анализе *in silico* всего генома путем моделирования виртуальной мультиплексной ПЦР с произвольными праймерами, а не отдельно выбранного фрагмента ДНК, что имеет неоспоримое преимущество, в том числе в виде гигантского числа комбинаций выявляемого полиморфизма, что может служить для идентификации таксонов более низкого уровня, чем это позволяет делать ДНК-баркодирование, что крайне важно для селекционных работ по созданию новых сортов растений и пород животных.

ПРИНЦИП ГЕНОМНОГО БАРКОДИРОВАНИЯ

В основе предлагаемого метода геномного баркодирования лежит перевод данных, полученных в результате ПЦР и капиллярного гель-электрофореза, в цифровой формат, а именно в бинарный (0, 1), который удобно представить в виде штрихкода. В этом случае выявляемые при гель-электрофорезе полосы (пики) ДНК отождествляет 1, а отсутствие ДНК – 0.

ПЦР – это реакция, в результате которой происходит многократная амплификация отдельного фрагмента ДНК [11]. Основными компонентами ПЦР являются: молекулы ДНК (объект исследования); фермент ДНК-полимераза, строящий новую комплементарную цепь ДНК по исходной матрице; праймеры (синтезированные химическим путем короткие олигонуклеотидные фрагменты, обычно длиной от 10 до 30 нуклеотидов), служащие затравкой для работы ДНК-полимеразы; дезоксинуклеозидтрифосфаты – дАТФ, дГТФ, дЦТФ, ТТФ, являющиеся «строительным материалом» при полимеризации ДНК; буферный раствор (среда для проведения реакции).

Один цикл ПЦР состоит из трех стадий: денатурация (разделение цепей ДНК при температуре около 95 °С); отжиг праймеров (присоединение праймеров к цепочкам ДНК, «ограничивающих» целевой фрагмент ДНК – ампликон) при температурах обычно от 40 до 60 °С, что зависит от особенностей праймера – его длины и GC-состава); элонгация («достраивание» цепи ДНК с помощью ДНК-полимеразы при 72 °С в направлении 5’→3’). В результате одного цикла образуется две молекулы ДНК из одной. Обычно проводится 25–40 циклов, в ходе которых амплификация идет экспоненциально.

Схема стандартной реакции ПЦР представлена на рисунке 1.

В классической ПЦР используется два праймера – так называемые прямой и обратный, имеющие разные последовательности и ограничивающие целевой фрагмент ДНК, что требует знания нуклеотидной последовательности амплифицируемого участка ДНК. Однако имеется целый ряд вариаций ПЦР, позволяющих вести амплификацию неких анонимных участков ДНК без знания их нуклеотидных последовательностей, что некоторое время назад нами рассмотрено достаточно подробно [12]. Один из таких методов носит название RAPD-анализа (Rapid Amplified Polymorphic DNA) [13] и его особенностью является применение только одного праймера, который выступает как прямым, так и обратным, что иногда вызывает удивление, которое здесь необходимо снять, используя рисунок 2. При этом нужно напомнить, что ДНК – двухцепочечная молекула, цепи в которой объединяются по принципу комплементарности и антипараллельны. Не вдаваясь в детали почему так, отметим, что один конец цепи ДНК, считающийся начальным, обозначается как 5’,

Математическая биология и биоинформатика. 2022. Т. 17. № 2. doi: 10.17537/2022.17.208

тогда как другой конец – 3’. Также важно указать, что в базах данных приводится только одна цепь ДНК – кодирующая РНК (когда это известно) и условно считающаяся «верхней».

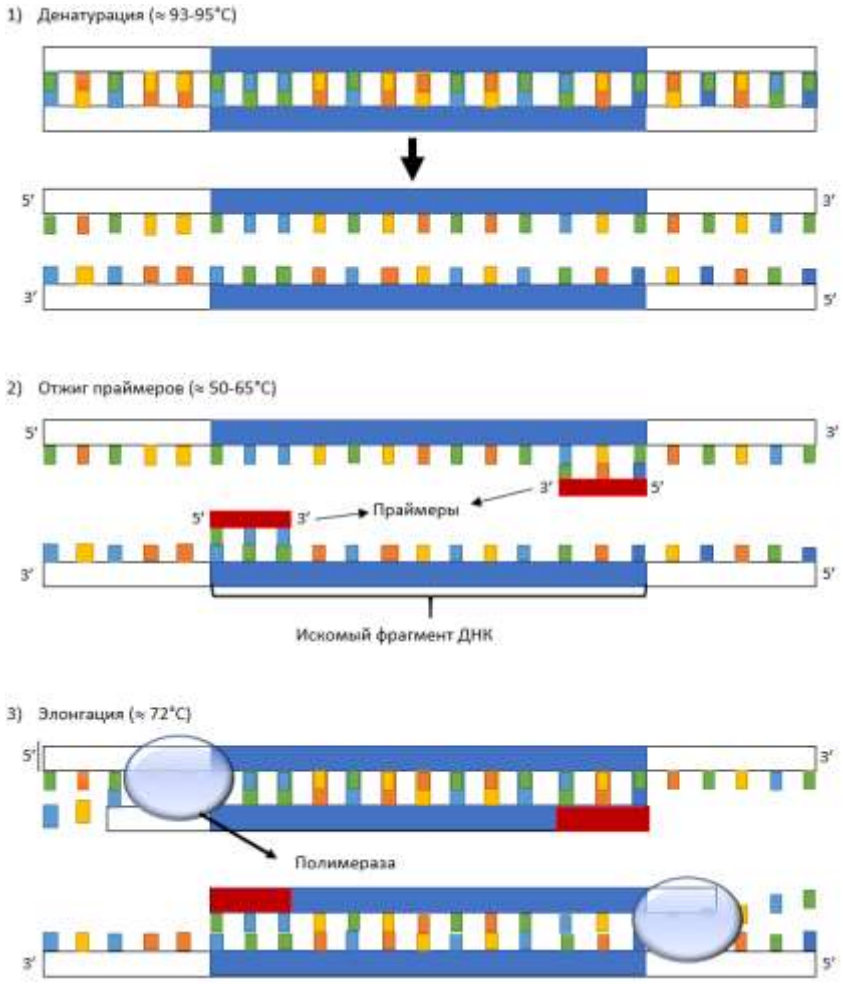


Рис. 1. Этапы проведения ПЦР (пояснения в тексте).

Как можно видеть из рисунка 2, в качестве прямого и обратного праймера выступает один и тот же олигонуклеотид с последовательностью 5’-AACCAGACAA-3’ и это происходит так потому, что на участке ДНК на разных цепях находятся комплементарные последовательности.

5′-NNNAACCAGACAANNN...NNNTTGTCTGGTTNNN-3′ *(верхняя цепь ДНК)*
3′-AACAGACCAA-5′ *(обратный праймер)*

5′-AACCAGACAA-3′ *(прямой праймер)*
3′-NNNTTGGTCTGTTNNN...NNNAACAGACCAANNN-5′ *(нижняя цепь ДНК)*

Рис. 2. Фрагмент двухцепочечного участка ДНК, превратившегося после этапа денатурации в две отдельные цепи, на которых отожглись праймеры 5’-AACCAGACAA-3’. Здесь N – любые нуклеотиды, а многоточие означает протяженную последовательность неопределенной длины. Остальные пояснения в тексте.

В отличие от стандартной ПЦР при проведении мультиплексной ПЦР используется набор праймеров, количество которых может меняться в довольно широких пределах и все зависит от стоящих перед экспериментатором задач и возможностей используемого метода. При этом праймеры в мультиплексной ПЦР, если они не нацелены на

определенные участки генов, что как раз имеет место в RAPD-анализе, образуют между собой пары во всех возможных сочетаниях, что увеличивает количество образующихся ампликонов разных размеров, что крайне важно для обеспечения выявления должного уровня полиморфизма ДНК. Максимально возможное количество пар праймеров, служащих прямыми и обратными, в этом случае оценивается как N^2, где N – количество праймеров. Принцип отжига праймеров при мультиплексном RAPD-анализе представлен на рисунке 3.

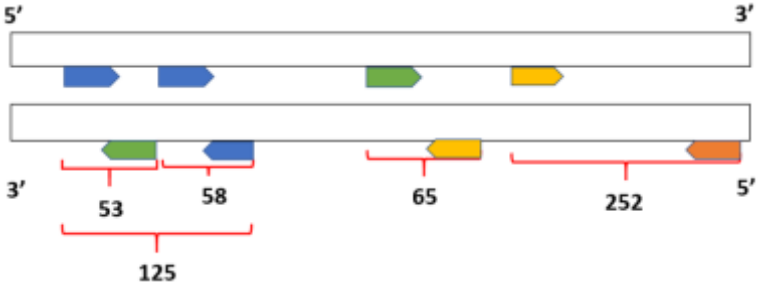


Рис. 3. Отжиг произвольных праймеров при мультиплексном RAPD-анализе. Разными цветами отмечены различные праймеры, числами обозначены размеры ампликонов в нуклеотидах при отжиге праймеров.

Для построения геномного штрихкода какого-либо организма нами предлагается использовать мультиплексный RAPD-анализ. При этом целью мультиплексного RAPD-анализа является увеличение числа ампликонов разного размера, часть которых могут оказаться полиморфными у близких видов растений (сортов), что позволит их различать между собой. Достижение нужного числа (полиморфных) ампликонов определяется как количеством праймеров в мультиплексной ПЦР, так и длиной самих праймеров, поскольку в зависимости от их длины они могут теоретически встречаться в геномах в среднем через некоторое количество нуклеотидов, составляющее для типичных RAPD-анализов декамерных праймеров приблизительно миллион нуклеотидов (4^{10}). Другой модификацией такого варианта RAPD-анализа является разделение ампликонов с помощью секвенирующего гель-электрофореза, в котором разделяются фрагменты ДНК, образовавшиеся с использованием флуоресцентно-меченных праймеров [14]. Продукты ПЦР предлагается разделять в виде их одноцепочечных вариантов в геле в денатурирующих условиях в генетическом анализаторе (капиллярный автоматический секвенатор). Данный прибор позволяет определять размеры ампликонов в некотором диапазоне с точностью до одного нуклеотида, тогда как при классическом RAPD-анализе продукты амплификации обычно разделяются в агарозном геле и оцениваются ампликоны, имеющие размеры от 200 до 2000 пар нуклеотидов (п.н.), что принципиально не позволяет определять их точные размеры, а допущение даже ±1 уже не позволяет считать такие данные истинно цифровыми.

В этой связи нужно заметить следующее. Биология никогда не относилась к точным наукам, поскольку в ней, если и происходят какие-либо измерения, например, размеров листьев и стеблей, то применяются большие допущения ввиду того, что подобные параметры могут весьма сильно отличаться даже у одного растения. Но в последние десятилетия ситуация несколько поменялась. В 1991 г. N.E.Morton свою статью, посвященную геному человека (размерам хромосом, определенных радиоавтографией и проточной цитофлуориметрией), начал со слов, что каждая наука имеет параметры точного определения чего-то важного для нее, и привел примеры: для астрономии – скорость света, а для химии – число Авогадро [15]. При этом он обратил внимание, что для молекулярной биологии таких параметров до сих пор вроде как нет, поскольку она пока пользуется, например, тем же числом Авогадро. Однако нам представляется, что для молекулярной биологии, по крайней мере, для части, связанной с нуклеиновыми

кислотами и геномами, важным параметром стало число нуклеотидов, определяемое некоторыми методами с точностью до одного азотистого основания, тем более что это прямой путь к удобной оцифровке ДНК-данных. Так, нуклеотид (нт) или их пара (п.н.) в виде звеньев соответственно одно- или двухцепочечной ДНК могут рассматриваться как важный (измерительный) параметр для молекулярной биологии и для анализа геномов, в частности, на чем и основано предлагаемое нами геномное баркодирование, для построения которого был выбран диапазон от 51 до 500 нуклеотидов. Установленные размеры ампликонов предлагается переводить в бинарную последовательность длиной 450 из неких ДНК-ячеек, где 1 ставится в случае наличия ампликона определенной длины, 0 – в случае его отсутствия. Фактически мы имеем шкалу от 51 до 500, на которой 1 ставится в случае наличия ампликона конкретной длины.

Принцип формирования геномного штрихкода представлен на рисунке 4.

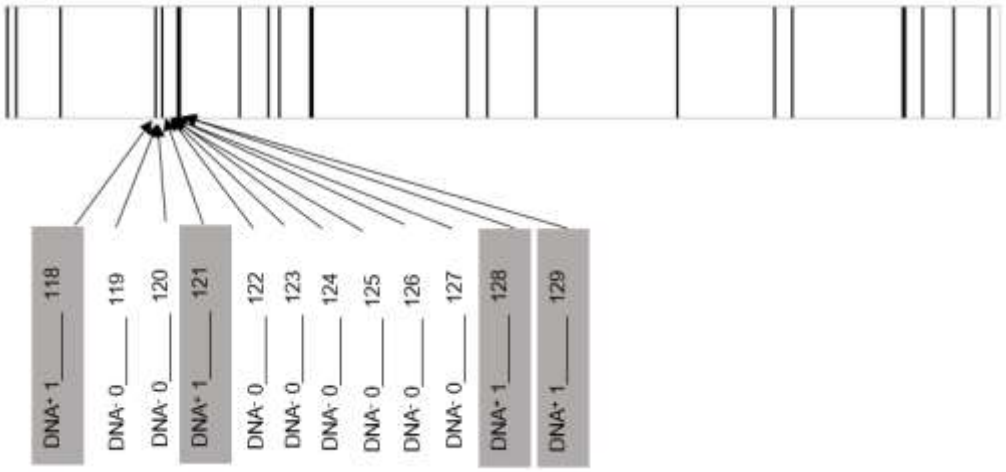


Рис. 4. Пример формирования геномного штрихкода.

Теоретическое число комбинаций встречаемости в этих воображаемых ДНК-ячейках ампликонов разного размера рассчитывается как число сочетаний без повторений из *amp* возможных ДНК-ячеек по amp_+ согласно формуле:

$$C_{amp}^{amp_+} = \frac{amp!}{amp_+!(amp - amp_+)!},$$

где C – общее число встречаемости комбинаций фрагментов ДНК выбранного размерного диапазона, *amp* – число всех анализируемых в выбранном диапазоне воображаемых ДНК-ячеек (450), amp_+ – число ДНК$^{(+)}$-ячеек, $(amp - amp_+)$ – число ДНК$^{(-)}$-ячеек. Максимальное число комбинаций, заметно превышающее гугол (более чем 10^{134}) геномных штрихкодов для выбранного нами для анализа диапазона фрагментов ДНК с размерами от 51 до 500 пар нуклеотидов, будет достигаться, если занятыми окажутся 225 ДНК$^{(+)}$-ячеек из 450 ДНК-ячеек. Однако и гораздо меньшее количество образующихся ампликонов (ДНК$^{(+)}$-ячеек) способно обеспечить уникальность геномных штрихкодов для разных организмов. Так, пример с 23 ампликонами на рисунке 4 обеспечивает около ундециллиона комбинаций их расположения в этом диапазоне.

Хотя выше говорилось, что использование RAPD-анализа не требует знания нуклеотидных последовательностей, информация о полном геноме организма позволяет провести виртуальную ПЦР *in silico* и спрогнозировать эффективность применения тех или иных праймеров и/или их комбинаций, что дает возможность без проведения экспериментов подобрать наиболее оптимальные из них и только потом их

синтезировать, поскольку это стоит немалых денег, тем более с учетом флуоресцентных меток, а также занимает больше времени.

К тому же подбор праймеров для проведения ПЦР особенно в ее мультиплексном варианте является довольно непростой задачей и вручную это часто не представляется возможным сделать. Для этого разработано большое количество программ, которые позволяют проводить дизайн праймеров согласно определенным требованиям эксперимента и их краткое рассмотрение было проведено нами некоторое время назад [16]. Однако, когда есть информация о полном геноме того или иного организма появляется возможность, прежде чем проводить RAPD-анализ в лабораторных условиях определить для подобранных праймеров предполагаемые места отжига, чтобы понять, можно ли в принципе получить желаемое количество ампликонов. Поэтому немаловажной задачей является этап планирования подобного эксперимента в виде RAPD-анализа, ведущего к формированию геномного штрихкода. Для этой цели нами была разработана компьютерная программа, которая позволяет прогнозировать возможные места отжига праймеров, определять необходимое количество праймеров для мультиплексного RAPD-анализа в «идеальных» условиях проведения ПЦР.

ПРОГРАММА ПОИСКА МЕСТ ОТЖИГА ПРАЙМЕРОВ ДЛЯ ГЕНОМНОГО БАРКОДИРОВАНИЯ

Для того чтобы проанализировать предполагаемый результат проведения ПЦР была разработана программа ABCDNA_GS, которая позволяет находить места отжига произвольных праймеров и определять размеры ограничиваемых ими фрагментов ДНК [17]. Стоит обратить внимание на сделанные нами допущения: в программе учитывается отжиг праймера по всей его длине, то есть не рассматриваются варианты, при которых какой-либо нуклеотид в праймерной последовательности останется неспаренным с мишенью в геноме, а также не учитывается возможное образование шпилечных структур во фланкирующих местах отжига участках геномной ДНК, при этом праймеры подбираются таким образом, что они ни при каких условиях не способны сформировать гомо- и гетеродимеры, в том числе в мультиплексном варианте, чему нами ранее уделено значительное внимание [18].

Фактически задача сводится к поиску позиций шаблонов в строке на определенном расстоянии друг от друга. Шаблонами являются праймеры, а строкой – исследуемый геном. Общий принцип работы программы представлен на рисунке 5.

Рис. 5. Общий принцип поиска позиций праймеров для геномного баркодирования.

Математическая биология и биоинформатика. 2022. Т. 17. № 2. doi: 10.17537/2022.17.208

Программа реализована на языке программирования Python 3.6 [19] с применением библиотеки Biopython. В качестве входных данных используются файлы, в которых хранятся нуклеотидные последовательности геномов в форматах *fasta, *.fa, *.fna, *.fastaq, а также праймеры (прямые), которые вводятся в виде строки через запятую. Геномы были взяты из специализированных баз данных GenBank [20], 1001Genomes [21], NIH [22].

Реализованы два формата представления информации о геноме: нуклеотидные последовательности хранятся в виде нескольких файлов, в каждом из которых хранится хромосома; нуклеотидные последовательности хранятся в виде общего файла, разделенного на фрагменты – контиги (отдельные несвязанные фрагменты ДНК) или цельные хромосомы. Так как в файле (как уже говорилось выше) хранится последовательность одной цепи ДНК для того, чтобы не восстанавливать вторую цепь ДНК, можно рассчитать позиции отжига обратных праймеров на заданной цепи. Для этого необходимо сформировать обратные праймеры, учитывая принцип комплементарности (A↔T, G↔C) и то, что цепи ДНК противоположно направлены. Пример формирования обратного праймера представлен ниже:

(5′)AACCAGACAA (3′)	Прямой праймер
↕↕↕↕↕↕↕↕↕↕	
(5′)TTGGTCTGTT (3′)	Комплементарная цепь в месте отжига прямого праймера
(3′)AACAGACCAA (5′)	Обратный праймер

Далее программа считывает последовательность генома и проводит в ней поиск вхождений каждого праймера (прямого и обратного). Стоит отметить, что для прямого праймера находится позиция расположения первого нуклеотида, а для обратного праймера – позиция последнего нуклеотида в цепи ДНК, при этом они совпадают с 5’-концом праймеров. Поиск праймера в геноме осуществляется с помощью алгоритма Кнута – Мориса – Пратта [23]. Если геном состоит из нескольких контигов, то поочередно считывается каждый контиг.

Далее формируются словари со следующей структурой:

Direct_positions = {‘Primer’: ‘нуклеотидная последовательность’, ‘Direction’: ‘direct’, ‘Positions’: [массив найденных позиций]}

Inverse_positions = {‘Primer’: ‘нуклеотидная последовательность’, ‘Direction’: ‘inverse’, ‘Positions’: [массив найденных позиций]}.

После чего для каждого обратного праймера формируется пара с прямым праймером, расстояние между которыми удовлетворяет диапазону [51; 500]. В выходном файле хранится информация в следующем виде: прямой праймер, позиция прямого праймера в геноме, обратный праймер, позиция обратного праймера в геноме, размер ампликона, номер контига/хромосомы. Стоит отметить, что разным наборам праймеров в рамках исследования одного генома будут соответствовать разные ампликоны и, как следствие, разные геномные штрихкоды.

Пример получаемых данных для генома резуховидки Таля *Arabidopsis thaliana* представлен в таблице 1.

Из таблицы 1 видно, что ожидаемое количество ампликонов для *A. thaliana*, имеющей небольшой геном размером всего около 135 млн. п.н., явно не обеспечит должного уровня полиморфизма, поэтому для мультиплексной ПЦР число используемых праймеров должно быть заметно увеличено и предварительно проверено путем проведения виртуальной ПЦР *in silico*.

Таблица 1. Результаты анализа генома *Arabidopsis thaliana*, набор праймеров: AACCAGACAA, AAGGGACAAA, AACCGAACAA, AACGCACAAA, AAAACGCCAA, AACGCCAAAA

Место отжига прямого праймера	Позиция 5'-нуклеотида прямого праймера	Место отжига обратного праймера	Позиция 5'-нуклеотида обратного праймера	Размер ампликона
AAGGGACAAA	29175626	TTGTTCGGTT	29176093	467
AACGCACAAA	12596171	TTTGTGCGTT	12596526	355
AACGCCAAAA	11447030	TTGGCGTTTT	11447142	112
AAGGGACAAA	10114337	TTGTCTGGTT	10114720	383
AAGGGACAAA	29175626	TTGTTCGGTT	29176093	467

РЕЗУЛЬТАТЫ КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ

Предложенный метод геномного баркодирования был апробирован на ряде родственных растений и животных. Ранее нами был проведен анализ семейства Крестоцветных на примере *A. thaliana* [24], который позволил заметить, что для более полного и наглядного анализа геномов родственных организмов важно выбрать достаточное количество праймеров, для того чтобы получить количество ампликонов, которые бы позволяли различать по штрихкодам анализируемые образцы. Здесь требуется акцентировать внимание на том, что нахождение в конкретной ДНК$^{(+)}$-ячейке одного или большего числа разных ампликонов, имеющих одинаковый размер, не принципиально, поскольку в этом методе в данном аспекте ведется качественный, а не количественный анализ, и такая ДНК-ячейка все равно должна оцифровываться как «1». К тому же определение в каждой такой ячейке истинного числа разных ампликонов с одинаковыми размерами довольно затруднительно, грешит неточностями и фактически не требуется для формирования геномного штрихкода. Вполне может быть, что у разных образцов окажутся случайно совпадающие по размеру ампликоны, не связанные между собой родством. Может быть и иная ситуация, когда ампликоны, отличающиеся между собой по размеру, могут принадлежать, по сути, одинаковым участкам ДНК и подобное на примере *A. thaliana* нами было обнаружено. Так, сопоставление нуклеотидных последовательностей ампликонов, отличающихся по размеру у разных образцов на один нуклеотид выявило, что они схожи между собой, но несут помимо трех однонуклеотидных замен, одну делецию. Однако данные ампликоны, несмотря на то, что они происходят из одинаковых участков генома и родственны для целей геномного баркодирования, к сожалению, одинаковыми считать нельзя, поскольку этот наш подход с геномным баркодированием на основе размеров ампликонов не рассчитан на установление нуклеотидных последовательностей всех или даже части ампликонов, так как за счет большого числа комбинаций решает другую задачу (относительно) быстрой ДНК-идентификации того или иного образца (сорта растений). При этом специально проведенное нами сравнение нуклеотидных последовательностей ряда ампликонов с одинаковыми размерами у линий *A. thaliana* с близкими геномными координатами показало, что они гомологичны друг другу на 100 % и именно на подобные фрагменты ДНК делается упор в геномном баркодировании по предлагаемому нами принципу. Здесь можно также заметить, что чем длиннее ампликон, тем с большей вероятностью он может по размеру отличаться от родственного ему за счет возможных инсерций и делеций, но выбранный нами диапазон длин от 51 до 500 нуклеотидов позволяет рассчитывать, что массовый характер инделы у них не носят, что видно, в том числе, при сравнении близкородственных организмов, например, у возделываемых видов пшениц и их диких сородичей. Но, прежде чем переходить к описанию геномного баркодирования последних, нужно уделить этим злакам чуть больше внимания.

Так, возделываемые твердая (*Triticum durum*) и мягкая (*T. aestivum*) пшеницы являются довольно сложными природными образованиями и состоят соответственно из двух и трех родственных субгеномов – B, A и B, A, D. То есть первый вид возник после скрещивания с диплоидной пшеницей (предположительно с *T. urartu*) диплоидного эгилопса *Aegilops speltoides* (также предположительно), после чего уже образовавшаяся тетраплоидная пшеница скрестилась с другим эгилопсом *Ae. tauschii* – носителем субгенома D в результате чего возникла гексаплоидная мягкая пшеница, ныне широко возделываемая. Причем она формально на две трети эгилопс, а не пшеница, но данные виды придумал человек и среди ботаников есть те, кто предлагает все эгилопсы и пшеницы признать относящимися к роду Triticum. Это несколько упрощенное представление филогении пшениц, но интерес к ней не праздный. Дело в том, что виды пшенично-эгилопсного альянса (которых существует около полусотни) представлены не менее чем 10 разнокачественными диплоидными геномами, а Природа при создании мягкой хлебной пшеницы использовала лишь три, причем два первых из них не совсем понятно от кого произошли, что крайне важно знать, поскольку у селекционеров имеется возможность создать новую искусственную пшеницу с превосходящими нынешнюю по целому ряду хозяйственно-полезных признаков, ввиду того, что Природа могла выбирать лишь из тех видов, что произрастали совместно. Более подробно ситуация с донорством субгеномов полиплоидных пшениц и их эволюцией рассмотрена нами ранее [25].

Для анализа геномов пшениц и эгилопсов с целью их геномного баркодирования был выбран следующий набор праймеров: AACCAGACAA, AAGGGACAAA, AACCGAACAA, AACGCACAAA, AAAACGCCAA, AACGCCAAAA, TGCCACACAC, AGCCTCACTC, TGCTCACCAC, CGACTCTCAC, AGCTCTCCAC, AGCTCCACTC.

Результаты полученных ампликонов представлены в таблице 2.

Таблица 2. Сводная таблица найденных ампликонов и сформированных геномных штрихкодов при компьютерном моделировании мультиплексной ПЦР для пшеницевых

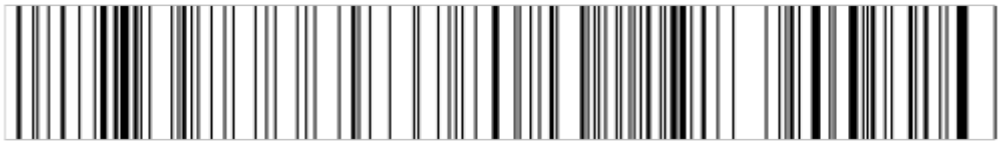
Разновидность	Ампликоны	Бинарная последовательность	Штрихкоды
Triticum aestivum **BAD** genome (около 15 млрд. п.н.)	051 052 053 055 056 057 060 061 063 066 067 068 072 073 074 075 076 077 078 080 081 082 084 086 088 091 093 094 095 096 098 100 102 103 104 105 106 109 110 112 113 116 118 120 121 122 124 127 128 129 130 131 132 133 134 135 136 137 138 144 150 154 157 158 162 163 164 165 167 169 171 172 173 177 179 180 184 185 187 188 191 194 195 196 198 201 202 208 209 211 212 216 221 222 225 226 227 228 229 235 236 237 238 241 242 243 244 245 246 248 250 252 254 255 256 257 258 259 260 263 264 267 268 271 272 273 274 275 277 279 280 282 283 284 288 289 290 292 293 294 296 297 298 299 301 302 304 307 313 314 315 316 318 320 322 323 324 325 326 328 329 330 331 335 336 337 338 339 341 342 343 344 345 346 347 348 349 350 352 353 354 355 356 357 358 359 360 361 362 363 365 367 368 369 369 370 372 373 374 375 377 378 379 380 381 382 383 384 386 387 388 392 394 395 396 397 398 399 402 404 405 406 407 408 409 414 415 416 417 418 420 421 426 427 428 430 435 436 437 438 439 440 441 444 447 448 449 450 452 453 456 457 458 461 462 463 464 465 466 467 468 469 470 471 472 474 475 476 478 479 480 482 483 484 485 486 487 488 489 491 493 493 494 495 496 497 498 499 500	11101110011010011100011 11111011101010100101111 01010111110011011001010 11101001111111111110000 01000001000100110001111 01010111000101100011011 00100111010011000001101 10001000011001111100000 11110011111101010101111 11100110011001111101011 01110001110111011110110 10010000011110101011111 01111000111110111111111 10111111111111010111101 11101111111101110001011 11110010111111000011111 01100001110100001111111 00100111101100111001111 11111111011101110111111 1101011111111	
Triticum aestivum subgenome **A**	052 053 056 057 063 066 076 077 084 086 091 094 095 096 098 100 103 104 105 106 109 110 112 116 122 129 130 131 135 138 144 150 154 164 167 169 173 177 187 188 191 195 198 201 202 208 209 211 212 222 225 236 238 242 244 252 258 263 272 273 274 275 277 282 284 288 289 293 297 299 301 314 315 318 320 323 328 336 337 339 343 344 348 352 353 355 356 357 358 359 360 365 367 374 381 386 394 398 402 405 407 417 418 421 428 435 436 437 439 440 450 456 461 462 464 468 469 474 476 478 483 484 485 486 487 491 498 500	01100110000010010000000 00110000001010000100111 01010011110011010001000 00100000011100010010000 01000001000100000000010 01010001000100000000011 00100010010011000001101 10000000001001000000000 01010001010000000100000 10000100000000111101000 01010001100010001010100 00000000001100101001000 01000000011010001100010 00110111111000010100000	

Математическая биология и биоинформатика. 2022. Т. 17. № 2. doi: 10.17537/2022.17.208

Разновидность	**Ампликоны**	**Бинарная последовательность**	**Штрихкоды**
		0100000010000010000000010 0010001001010000000011 0010000001000000110110 0000000010000010000101 0001100001010100001111 0001000000101	
Triticum aestivum subgenome **B**	051 061 067 068 072 074 076 078 081 082 084 091 093 094 095 098 103 105 113 118 122 127 132 133 134 136 137 154 157 158 162 165 171 172 179 180 184 191 194 196 209 221 228 229 237 238 243 245 246 248 250 254 255 256 257 258 264 267 271 273 277 283 288 288 294 297 298 302 304 307 313 320 326 326 335 337 338 342 344 345 346 348 349 350 353 355 356 357 358 359 360 361 362 363 368 369 373 380 381 382 383 384 395 396 397 399 402 404 406 414 416 427 437 439 440 444 447 449 452 453 458 461 463 464 466 468 469 470 471 472 476 478 480 482 484 486 487 488 491 495 496	1000000000100000110001 0101001101000000101110 0100001010000001000010 0010000100001110110000 0000000000100110001001 0000011000001100010000 0010010100000000000100 0000000001000000110000 0011000010110101000111 1000001001000101000001000 0010000100000100110001 1001000001000000100001 0000000010110001011101 1001011111111100001100 1000000111110000000001 1101001010100000010100 0000000010000000010110 0010010100110000100101 0101111100010101010101 1001000110000	
Triticum aestivum subgenome **D**	051 055 060 061 063 072 073 074 075 080 088 093 094 096 102 105 118 120 121 124 128 130 131 154 163 169 173 185 188 202 209 212 216 226 227 228 235 241 243 244 246 256 258 259 260 268 272 277 279 280 288 290 292 296 299 301 302 316 322 323 324 325 329 330 331 341 347 352 353 354 357 358 362 363 367 369 370 372 374 375 377 378 379 387 388 392 394 396 397 399 402 406 408 409 415 420 426 430 437 438 439 441 447 448 456 457 458 465 467 469 474 475 479 483 486 489 491 493 494 497 498 499	1000100001101000000011 1100001000000100001101 0000010010000000000010 1100100010110000000000 0000000000010000000100 0001000100000000001001 0000000000000010000000100 1000100000000111000000 1000001011010000000010 1110000000100010001011 0000000101010001010110 0000000000001000011110 0011100000000100000100 0011100110001100010 1101 0110111000000110001010 1101001000101100000 1000 0100000100010000011101 0000011000000111000000 1010100001100010001 0010 0101011001110	

Разновидность	Ампликоны	Бинарная последовательность	Штрихкоды
T. turgidum **BA** genome (около 10 млрд. п.н.)	051 053 057 061 063 066 067 068 072 074 076 077 083 084 085 091 093 094 095 096 098 100 101 103 104 105 106 109 110 111 112 113 116 118 122 127 128 129 130 131 132 133 134 135 137 138 144 150 152 154 158 160 162 163 164 165 167 169 171 172 173 177 180 183 184 187 189 191 194 195 197 198 201 202 203 208 209 211 221 225 228 229 236 237 238 242 243 244 245 246 247 248 250 252 255 256 257 258 259 264 265 267 270 271 272 273 274 275 277 278 279 282 283 284 287 288 289 293 294 297 298 299 301 302 304 312 313 314 318 319 320 323 326 327 328 332 335 336 337 338 339 343 344 345 346 347 348 349 350 352 353 355 356 357 358 359 360 361 362 363 365 366 367 368 370 373 374 379 380 381 382 384 385 387 395 396 397 402 405 406 407 414 416 417 418 424 429 431 433 435 436 437 439 440 441 444 445 447 451 452 456 458 461 462 463 464 465 468 469 470 471 472 474 475 476 477 478 481 482 483 484 485 486 487 488 489 491 493 498 500	1010001000101001110001010110001000011100000101111 0101101111001111100101000100001111111110110000 0100000101010001010111101010110001001100010010 0110001111111101010011110000110100111111101110 0111001110001100111011000000000100010011000001 1000000011100011100100111000011110100111 1100010011111100011111111011011111111110111010000 1100001111011010000000111000011101000000010111 0000010000101010111011100011010001100101001111 1001111101111100111111111101010000101	
T. turgidum subgenome **A**	053 057 063 066 067 076 077 084 091 094 095 096 098 100 103 104 105 106 109 110 112 116 122 127 128 129 130 131 132 135 138 144 150 154 164 167 169 173 177 187 189 191 195 198 201 202 208 209 211 225 236 238 242 244 247 252 258 264 272 273 274 275 278 279 282 284 287 288 289 293 298 299 301 312 313 314 318 320 323 327 328 332 336 337 339 343 348 350 352 355 356 357 358 359 360 367 368 374 381 385 396 402 405 407 417 418 424 433 435 436 437 439 440 441 456 461 462 464 465 468 469 472 474 477 478 483 484 486 487 491 498 500	0010001000001001100000000110000100111 0101001111001101000100000010000100111 0010000111111001001000000101000010000 0100000100010000000001000000000010 0101000100010000000001000000000010 1010001001001100000110100000000000001000000000 0101000101001000010000000010000110110 1000001000000011110011001010001100010 0101001110001000011010000000111000101001000 1100010001101000100001000000011110000 1010011111100000011000001000000000000 1000001001010000000001100000000000101110111 0000000000000001000011010011001010011000011011 0001000000101	

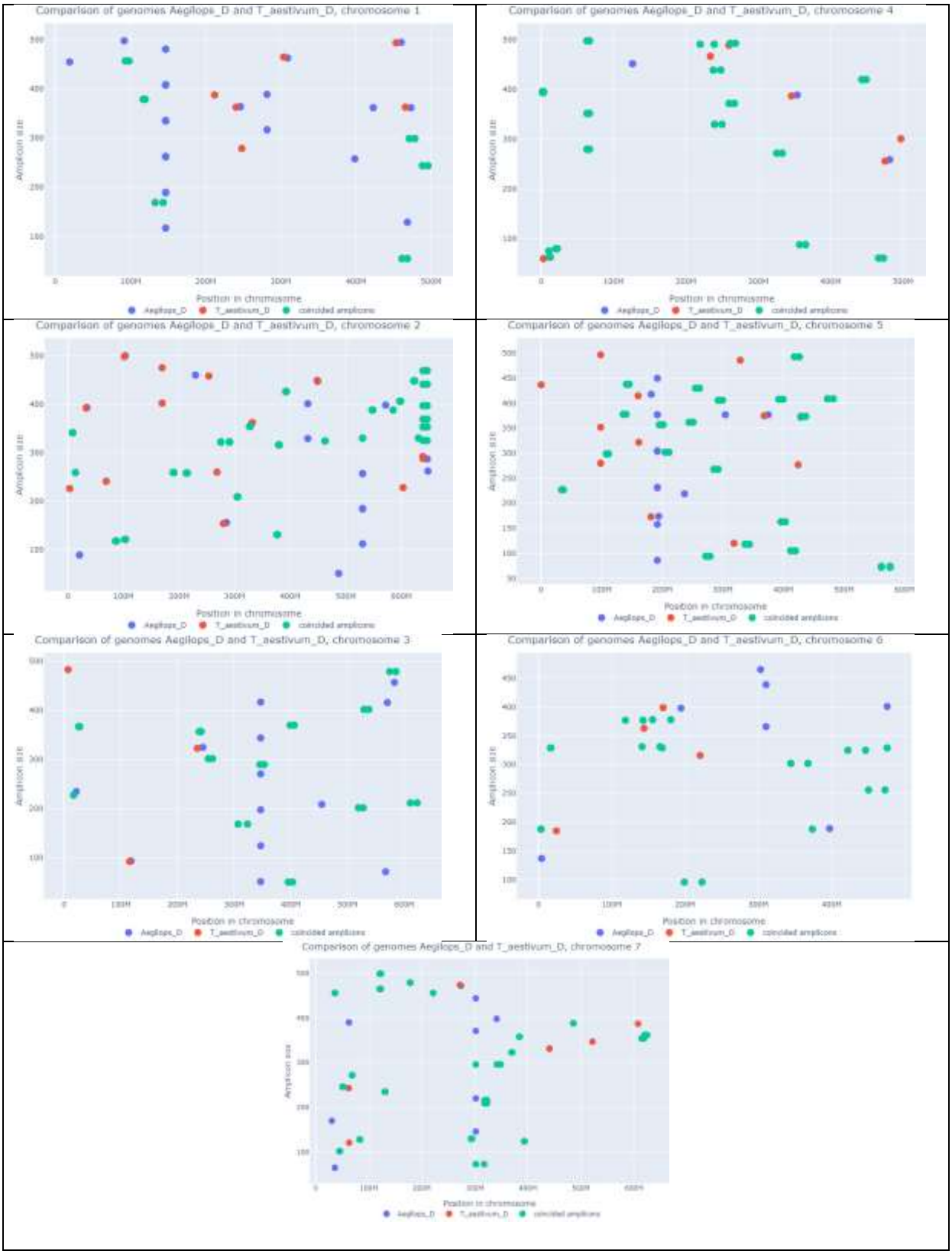
Разновидность	Ампликоны	Бинарная последовательность	Штрихкоды
T. turgidum subgenome **B**	051 061 067 068 072 074 076 083 085 091 093 094 095 098 100 101 105 111 113 118 122 127 130 132 133 134 137 152 158 160 162 163 165 171 172 180 183 184 191 194 197 203 209 221 228 229 237 243 245 246 248 250 255 256 257 258 259 265 267 270 271 277 283 288 294 297 298 302 304 318 319 320 326 332 335 337 338 344 345 346 347 349 352 353 355 356 357 358 359 360 361 362 363 365 366 370 373 379 380 381 382 384 387 395 397 402 406 414 416 418 429 431 436 437 439 440 444 445 447 451 452 458 463 468 469 470 471 472 475 476 478 481 482 485 487 488 489 491 493	100000000010000001100010 101000000101000000101110 010110001000001010000010 001000001001011100100000 000000000100000010101101 000001100000000100110000 001001001000001000000100 000000000100000001100000 001000000101101010000111 110000001010011000001000 001000001000001001100010 100000000000000111000001 000001001011000000111101 001101111111111011000100 100000111101001000000001 010000100010000000010101 000000000010100001101110 001101000110000001000010 00011111001101001100101 1101010000000	
T. dicoccoides **BA** genome (около 10 млрд. п.н.)	051 056 057 063 065 067 068 070 071 072 074 076 077 078 083 084 091 092 093 094 095 096 098 100 101 103 104 105 106 109 110 112 113 116 118 122 123 126 127 129 130 131 132 133 134 135 137 138 144 150 154 157 158 160 162 164 165 169 172 173 180 181 183 184 187 191 194 197 201 202 203 208 209 211 216 221 224 225 228 229 232 236 237 238 244 247 248 249 250 251 252 253 254 255 256 257 258 259 264 267 270 271 272 273 274 275 277 282 284 288 289 293 294 297 298 299 301 302 304 307 312 313 315 318 320 323 328 330 333 334 336 337 339 340 342 343 344 346 348 350 350 353 354 355 356 357 358 359 360 361 362 363 364 367 368 371 373 374 380 381 382 384 391 395 396 402 405 407 408 411 414 416 417 418 419 420 423 425 427 434 435 436 437 439 440 442 443 444 445 446 447 448 449 450 453 458 461 462 463 464 466 468 469 470 471 472 475 476 478 479 481 482 483 484 485 486 487 488 489 491 498 500	100001100000101011011110 101110000110000000111111 010110111100110110010100 001100110111111110110000 010000001000100110101011 000100110000000110110010 001001001000111000011101 000010000010011001100100 011100000100111111111111 110000100100111111101000 010100011000110011101100 100100000110100101001000 010100110110110111010100 100111111111111100110010 110000011101000000010001 100000100101100100101110 110010101000000111101100 111111111001000010011110 010111110011011011111110 1101000000101	

Разновидность	Ампликоны	Бинарная последовательность	Штрихкоды
T. dicoccoides subgenome **A**	056 057 063 065 070 076 077 084 091 094 095 096 100 101 103 104 105 106 109 110 112 116 126 129 131 132 135 138 144 150 154 164 169 173 183 187 191 202 208 209 211 216 225 236 238 247 252 255 258 264 272 273 274 282 284 284 289 293 298 299 301 312 313 315 318 320 323 328 330 334 336 339 342 343 348 350 353 354 355 357 358 359 362 367 368 374 381 396 402 405 407 408 411 417 418 419 420 425 427 434 435 436 437 440 442 444 445 450 453 461 462 464 468 469 475 478 483 484 485 486 487 500	000001100000010100001000 001100000010000000100111 000110111100011010001000 000000100101100100100000 010000010001000000000010 000100010000000000100010 001000000000001000001101 000010000000001000000000 010100000000010000100100 100000100000000111000000 010100001000100001100100 000000001101001010010000 010100010100100110000010 100111011100100001100000 010000001000000000000000 100000100101100100000011 110000101000000011110010 101100001001000000001101 000110000010010000111111 0000000000001	
T. dicoccoides subgenome **B**	051 067 068 071 072 074 076 078 083 091 092 093 094 095 098 100 103 105 113 118 122 123 127 130 132 133 134 135 137 157 158 160 162 165 172 180 181 184 194 197 201 203 209 221 224 228 229 232 237 238 244 248 249 250 251 253 254 256 257 258 259 264 267 270 271 273 275 277 282 288 294 297 298 302 304 307 313 318 320 330 333 337 340 344 346 348 350 353 354 355 356 358 360 361 362 363 364 371 373 374 380 381 382 384 391 395 402 414 416 417 423 439 443 444 446 447 448 449 458 463 466 468 469 470 471 472 475 476 478 479 481 482 487 488 489 491 498	100000000000000001100110 101010000100000000111110 010100101000000001000010 001100010010111110100000 000000000000000110101001 000000100000000110010000 000001001000101000000100 000000000100100011001000 001100000100011110111011 110000100100110101010000 010000010000010011000010 100100000100000101000000 000100100010010001010010 100111101011111000000010 110000011101000000010001 000000100000000000010110 000010000000000000000100 011011110000000001000010 010111110011011011000001 1101000000100	
Aegilops tauschii **D** genome (около 4,2 млрд. п.н.)	051 052 055 061 063 065 072 073 074 075 080 086 088 089 094 096 102 105 112 117 118 121 124 125 128 129 130 131 137 146 156 158 163 169 170 174 184 185 188 189 190 198 202 209 212 216 219 220 227 228 231 232 235 244 246 256 257 258 259 262 263 268 271 272 280 287 290 296 299 302	110010000010101000000011 110000100000010110000101 000001001000000010000110 010011001111000001000000 000100000000001010000100 000110001000000000011001 110000000100010000000100	

Разновидность	**Ампликоны**	**Бинарная последовательность**	**Штрихкоды**
	304 305 316 317 322 323 324 325 329 330 331 332 335 336 341 344 352 353 354 357 358 360 362 364 366 367 369 370 371 372 374 377 378 379 388 389 390 393 394 396 397 398 401 402 406 408 409 416 417 418 420 426 430 438 439 441 444 448 449 450 452 455 456 457 460 463 465 469 472 479 481 491 493 495 498 499 500	1000100110000001100110 1000000001010000000011 1100110000100110000001 0000001001000001001001 1100000000001100001110 0011110011000010010000 0011100110101010110111 0100111000000001110011 1110011000101100000111 0100000100010000001101 0010001110100111001001 1000100100000010100000 0001010100111	
T. urartu **A** genome (около 5 млрд.п.н.)	054 057 065 066 068 070 073 074 076 077 082 088 091 094 098 105 110 112 127 128 129 130 131 132 134 144 148 154 160 164 165 169 173 177 184 185 187 190 206 209 211 214 221 227 239 246 247 252 256 258 268 272 273 274 277 279 282 284 288 291 293 298 299 303 310 312 314 315 318 328 330 335 336 339 343 344 347 348 350 353 354 355 357 358 359 366 369 376 379 381 393 394 396 397 402 407 408 418 422 430 431 437 440 442 443 450 451 454 461 462 464 467 469 474 478 482 487 499	0001001000000110101001 1011000010000100100100 0100000010000101000000 0000000111111010000000 0100010000100000100011 0001000100010000011010 0100000000000000100101 0010000001000001000000 0000100000011000100010 1000000000100011001010 0101000100101000110001 0000001010110010000000 0101000011001000110110 1001110111000000101000 0001001010000000000110 1100001000011000000001 0001000000011000010010 1100000011001000001101 0010100001000100100001 0000000000010	

Как можно видеть из приведенных выше результатов, в зависимости от размеров геномов для одинакового комплекта мультиплексных произвольных праймеров меняется количество образующихся ампликонов.

Таблица 3. Сравнительный анализ геномов (по-хромосомно) на примере генома D *Aegilops tauschii* и субгенома D *Triticum aestivum*



Поскольку пшеницы и эгилопсы близкородственные виды и их геномы характеризуются высокой коллинеарностью и синтенией (за исключением ряда установленных транслокаций), то представляло интерес установление расположения одинаковых или близких по размеру ампликонов внутри разных субгеномов и геномов. Примеры такого сравнения представлены в таблице 3.

Аналогично данным таблицы 3 можно наблюдать сходства и различия между разными видами пшениц на уровне генома в целом и на уровне субгеномов. Но стоит отметить, что по полученным графикам при более глубоком анализе схожих по размеру ампликонов по их нуклеотидному составу следует обратить внимание на 2 группы ампликонов (пример на рис. 6).

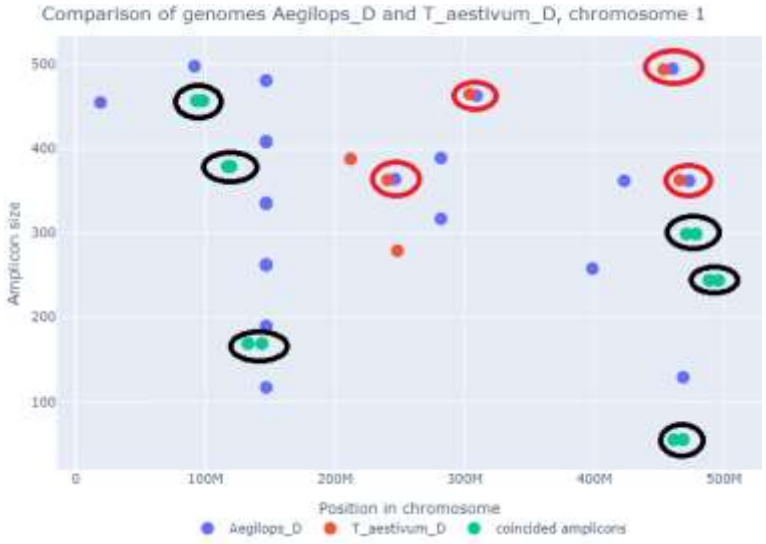


Рис. 6. Наглядное представление групп ампликонов, полученных в результате попарного компьютерного сравнительного анализа геномов пшеницевых.

Первая группа на рисунке 6 отмечена черным цветом. Это ампликоны, совпадающие по размеру и расположенные примерно на одинаковых позициях внутри хромосом у двух геномов/субгеномов. То есть, предполагается, что с высокой вероятностью это одинаковые по нуклеотидному составу фрагменты ДНК, обеспечивающие «общие черты» родственных организмов. Однако, не во всех случаях присутствие одинаковых по размеру фрагментов ДНК гарантирует их полное совпадение. Часть из них имеют незначительные различия с разницей до нуклеотида, что может говорить о единичных заменах нуклеотидов. Вторая группа выделена красным цветом. Это расположенные в геномах с учетом их коллинеарности поблизости друг от друга ампликоны, незначительно отличающиеся по размеру. Возможно, эти участки близки по нуклеотидному составу, но несут в себе различия на уровне нескольких нуклеотидов, что, впрочем, зная их локализацию в геноме, относительно несложно проверить, но предназначение геномного баркодирования несколько иное.

Помимо растений, описываемый метод баркодирования был применен нами для пород собак и лошадей [26, 27].

В целом, предложенный метод геномного баркодирования и компьютерного моделирования отжига праймеров мультиплексной ПЦР позволяет проводить компьютерный анализ родственных геномов, обнаруживать различия между ними еще до проведения натурного эксперимента. Данная методика позволяет формировать геномный штрихкод на основе анализа всего генома в целом, а не отдельного участка ДНК, что снимает проблему выбора эталонного фрагмента для классификации разных групп живых организмов.

ЗАКЛЮЧЕНИЕ

Здесь представлен новый метод компьютерного анализа родственных геномов. Разработана программа, позволяющая находить предполагаемые места отжига праймеров на цепях ДНК, и соответствующие им ампликоны. Предложен метод, который позволяет переводить данные, полученные в результате ПЦР в бинарный формат и в формат геномного штрихкода. Новый метод баркодирования позволяет проводить систематизацию организмов на уровне генома в целом, а не отдельно взятого фрагмента. Разработанная программа будет полезна на практике для биологов, генетиков, занимающихся классификацией и систематизацией живых организмов. А также это хороший вспомогательный инструмент для предварительного планирования экспериментов по выявлению мультилокусного полиморфизма ДНК. Если не проводить подобные анализы по уже известным геномам, то может оказаться, что подобранные праймер(ы) придутся на повторяющиеся элементы генома и тогда в «мокром» эксперименте будет образовываться чрезмерно большое количество разнообразных ампликонов, что будет отрицательно влиять на сбор первичных данных. Чтобы исключить такую возможность, требуется предварительный *in silico* анализ полных геномов исследуемых видов. Например, при проведенном нами анализе *in silico* генома бразильской гевеи, один из праймеров показал излишне много мест отжига и был в итоге забракован.

Также следует заметить, что нельзя исключать случаи, когда, например, произойдет буквально единичная мутация, которая повлечет за собой коренное изменение свойств сорта, выявляемое на фенотипическом или биохимическом уровне, но невидимое при избирательном анализе ДНК этого растения, поскольку замена такого нуклеотида окажется вне зоны отжига праймеров и, следовательно, будет пропущена. Однако и все остальные ныне используемые методы выявления полиморфизма ДНК с тем же «успехом» пропустят подобную мутацию, поскольку вероятность, что она окажется в анализируемом месте ничтожна. Только секвенирование всего генома теоретически может позволить такую мутацию найти. Но ДНК-паспортизация сортов и их ДНК-идентификация путем полногеномного секвенирования, несмотря на его бурное развитие и удешевление [28] вряд ли в ближайшие годы станет легко доступной, дешевой и уместной, хотя бы потому что для большинства целей достаточно охарактеризовать часть генома. Лишь при наличии очень серьезной необходимости выявления всех возможных отличий (мутаций) можно (нужно) будет прибегать (в будущем) к секвенированию всего генома. Здесь стоит провести аналогию с фотографией человека в его паспорте, которая не дает всего представления о нем как личности, но для целей первичной идентификации вполне пригодна.

Наконец стоит отметить, что построенный *in silico* геномный штрихкод даже при проведении RAPD-анализа с тем же образцом, что и был использован для полногеномного секвенирования не даст ту же исчерченность, поскольку в базах данных хранятся по сути квазигаплоидные геномы (включая их субгеномы), а в «мокром» эксперименте приходится иметь дело с двойными наборами хромосом, в реальности несколько отличающимися по последовательности от представленных в общемировых базах данных. Когда настанет время секвенирования полноценных диплоидных геномов, или двойного набора хромосом, покажет будущее, и этим вопросом мы задались в столетний юбилей самого термина «геном» [29]. Хотя и сейчас с помощью некоторых программ-сборщиков геномов и прочих ухищрений в виде дополнительных экспериментальных работ удается с не самой высокой точностью восстанавливать диплоидные геномы отдельных организмов, тем не менее желательно появление принципиально новой технологии секвенирования ДНК или существующие, вкупе с биоинформатической сборкой черновых данных, должны быть усовершенствованы настолько, что позволят секвенировать полные диплоидные

геномы практически любых организмов уверенно и массово. Однако, несмотря на чрезвычайную сложность этой задачи, нужно надеяться, что рано или поздно это произойдет, но возможно, стоит уже сейчас, спустя 100 лет после появления «генома», обозначающего по задумке автора H.Winkler гаплоидный набор хромосом, вводить в оборот обозначающий уже диплоидные геномы новый термин, например «дигеном» или более кратко «дином» (dinome), поскольку со временем секвенирование геномов в таком статусе станет необходимым и даже обязательным, по крайней мере для медицинской генетики. Но и для таксономии, и для селекционных работ знания полных диплоидных геномов окажутся крайне важны. Однако возвращаясь к нынешним (квази)гаплоидным геномам, необходимо отметить, что в настоящее время они приносят несомненную пользу, причем двояко – прогнозируя эффективность построения геномных штрихкодов в «мокром» эксперименте, на которые можно смело ориентироваться, а также позволяя устанавливать по ним родство исследуемых видов, благодаря гигантскому числу комбинаций ампликонов.

Исследование поддержано грантом Минобрнауки РФ (соглашение № 075-15-2021-1066 от 28 сентября 2021 г).

СПИСОК ЛИТЕРАТУРЫ

1. Болотова Н.Л. Биологическое разнообразие и проблемы его сохранения. В: *Наука – школе. Сборник научных публикаций*. 2017. С. 119–174.
2. Hebert P.D.N., Cywinska A., Ball S.L., deWaard J.R. Biological identifications through DNA barcodes. *Proceedings of the Royal Society*. 2003. V. 270. № 1512. P. 313–321. doi: 10.1098/rspb.2002.2218
3. Hebert P.D.N., Gregory T.R. The Promise of DNA Barcoding for Taxonomy. *Systematic. Biology*. 2005. V. 54. № 5. P. 852–859.
4. Ballard J.W.O., Whitlock M.C. The incomplete natural history of mitochondria. *Molecular. Ecology*. 2004. V. 13. № 4. P. 729–744. doi: 10.1046/j.1365-294x.2003.02063.x
5. Schoch C., Seifert K., Huhndorf S., Vincent R., Spouge J., Levesque A., Wen C. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*. 2012. V. 109. № 16. P. 6241–6246. doi: 10.1073/pnas.1117018109
6. Hollingsworth P. Forrest L. Spouge J., Hajibabaei M., Ratnasingham S., van der Bank M., Chase M., Cowan R. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*. 2009. V. 106. № 31. P. 12794–12797.
7. Nehal N., Choudhary B., Nagpure A., Gupta R.K. DNA barcoding: a modern age tool for detection of adulteration in food. *Critical Reviews in Biotechnology*. 2021. V. 41. № 5. P. 767–791. doi: 10.1080/07388551.2021.1874279
8. Chatpiyaphat K., Sumruayphol S., Dujardin J.-P., Samung Y., Phayakkaphon A., Cui L., Ruangsittichai J., Sungvornyothin S., Sattabongkot J., Sriwichai P. Geometric morphometrics to distinguish the cryptic species *Anopheles minimus* and *An. harrisoni* in malaria hot spot villages, western Thailand. *Medical and Veterinary Entomology*. 2021. V. 35. № 3. P. 293–301. doi: 10.1111/mve.12493
9. Panyukov V.V., Kiselev S.S., Alikina O.V., Nazipova N.N., Ozoline O.N. Short unique sequences in bacterial genomes as strain- and species-specific signatures. *Mathematical Biology and Bioinforatics*. 2017. V. 12. № 2. P. 547–558. doi: 10.17537/2017.12.547
10. Panyukov V., Kiselev S., Ozoline O. Unique k-mers as strain-specific barcodes for phylogenetic analysis and natural microbiome profiling. *Int. J. Mol. Sci*. 2020. V. 21. № 3. P. 944. doi: 10.3390/ijms21030944
11. Bartlett J.M.S., Stirling D. A Short History of the Polymerase Chain Reaction. *PCR Protocols*. 2003. V. 226. P. 3–6.

12. Кулуев Б.Р., Баймиев Ан.Х., Геращенков Г.А., Чемерис Д.А., Зубов В.В., Кулуев А.Р., Баймиев Ал.Х., Чемерис А.В. Методы ПЦР для выявления мультилокусного полиморфизма ДНК у эукариот, основанные на случайном праймировании. *Генетика*. 2018. Т. 54. С. 495–511. doi: 10.7868/S0016675818050016
13. Williams J.G., Kubelik A.R., Livak K.J., Rafalski J.A., Tingey S.V. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res*. 1990. V. 18. P. 6531–6535.
14. Corley-Smith G.E., Lim C.J., Kalmar G.B., Brandhorst B.P. Efficient Detection of DNA Polymorphisms by Fluorescent RAPD Analysis. *BioTechniques.* 1997. V. 22. № 4. P. 690–699. doi: 10.2144/97224st04
15. Morton N.E. Parameters of the human genome. *Proc. Natl. Acad. Sci.* 1991. V. 88. P. 7474–7476.
16. Чемерис Д.А., Кирьянова О.Ю., Губайдуллин И.М., Чемерис А.В. Дизайн праймеров для полимеразной цепной реакции (краткий обзор компьютерных программ и баз данных). *Биомика*. 2016. Т. 8. № 3. С. 215–238.
17. Кирьянова О.Ю., Кирьянов И.И., Кулуев Б.Р., Чемерис А.В., Гарафутдинов Р.Р., Губайдуллин И.М. *ABCDNA_GS* (*AMPLIFIED BAR-CODED DNA GENOME/SPECIMEN*): свидетельство о регистрации программы для ЭВМ № 2020610703 от 17.01.2020.
18. Гарафутдинов Р.Р., Баймиев А.Х., Малеев Г.В., Алексеев Я.И., Зубов В.В., Чемерис Д.А., Кирьянова О.Ю., Губайдуллин И.М., Матниязов Р.Т., Сахабутдинова А.Р., Никоноров Ю.М., Кулуев Б.Р., Баймиев А.Х., Чемерис А.В. Разнообразие праймеров для ПЦР и принципы их подбора. *Биомика*. 2019. Т. 11. № 1. С. 23–70.
19. *Python*. URL: https://www.python.org/downloads/release/python-360/ (дата обращения 15.09.2022).
20. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L. GenBank. *Nucleic Acids Res*. 2007. V. 36. P. D25–D30. doi: 10.1093/nar/gkm929
21. *1001Genomes*. URL: https://1001genomes.org/data-center.html (дата обращения 15.09.2022).
22. *National Library of Medicine*. URL: https://www.ncbi.nlm.nih.gov/ (дата обращения 15.09.2022).
23. Knuth D., Morris J.H., Pratt V. Fast pattern matching in strings. *SIAM Journal on Computing*. 1977. V. 6. № 2. P. 323–350. doi: 10.1137/0206024
24. Кирьянова О.Ю., Кулуев Б.Р., Кулуев А.Р., Марданшин И.С., Губайдуллин И.М., Чемерис А.В. Мультиплексный *in silico* RAPD-анализ ряда родственных растений с отличающимися размерами геномов и перспективы такого подхода для ДНК-паспортизации сортов сельскохозяйственных растений. *Биомика*. 2020. Т. 12. № 2. С. 194–210. doi: 10.31301/2221-6197.bmcs.2020-10
25. Кулуев А.Р., Матниязов Р.Т., Чемерис Д.А., Чемерис А.В. Современные представления о родственных взаимоотношениях в пшенично-эгилопсном альянсе (с краткой исторической справкой). *Биомика*. 2016. Т. 8. № 4. С. 297–310.
26. Кирьянова О.Ю., Гарафутдинов Р.Р., Чемерис Д.А., Гиниятов Ю.Р., Губайдуллин И.М., Чемерис А.В. Полиморфизм ДНК собак (*canis familiaris l.*). II. RAPD-анализ. *Биомика*. 2021. Т. 13. № 3. С. 309–320. doi: 10.31301/2221-6197.bmcs.2021-22
27. Гарафутдинов Р.Р., Гайнуллина К.П., Кирьянова О.Ю., Юрина А.В., Долматова И.Ю., Логинов О.Н., Чемерис А.В. Полиморфизм ДНК лошади *Equus Caballus* и методы его выявления. *Биомика*. 2020. Т. 12. № 2. С. 272–299. doi: 10.31301/2221-6197.bmcs.2020-16

28. Зубов В.В., Чемерис Д.А., Василов Р.Г., Курочкин В.Е., Алексеев Я.И. Краткая история методов высокопроизводительного секвенирования нуклеиновых кислот. *Биомика*. 2021. Т.13. № 1. С. 27–46. doi:10.31301/2221-6197.bmcs.2021-4
29. Кулуев Б.Р., Баймиев Ан.Х., Геращенков Г.А., Юнусбаев У.Б., Гарафутдинов Р.Р., Алексеев Я.И., Баймиев Ал.Х., Чемерис А.В. Сто лет гаплоидным геномам. Сейчас наступает время диплоидных. *Биомика*. 2020. Т. 12. № 4. С. 411–434. doi: 10.31301/2221-6197.bmcs.2020-33

Рукопись поступила в редакцию 08.07.2022, переработанный вариант поступил 18.08.2022.
Дата опубликования 27.09.2022.

========== **BIOINFORMATICS** ==========

Multiplex *in silico* RAPD-Analysis for Genome Barcoding

Kiryanova O.Yu.[1], Kiryanov I.I.[2], Kuluev B.R.[3], Garafutdinov R.R.[3], Chemeris A.V.[3], Gubaydullin I.M.[1,4]

[1]*Ufa State Petroleum Technological University, Ufa, Russia*
[2]*OOO "Corning SNG", Saint Petersburg, Russia*
[3]*Institute of Biochemistry and Genetics, Ufa Federal Research Center, Russian Academy of Sciences, Ufa, Russia*
[4]*Institute of Petrochemistry and Catalysis of Russian Academy of Sciences, Ufa, Russia*

Abstract. In this work, we propose a new method for identifying organisms of multiplex polymerase chain reaction (PCR) with arbitrary primers *in silico* (multiplex *in silico* RAPD-analysis) for the unique identification of living organisms. The results of computer modeling search of possible primer annealing sites in genomic DNA, and their convertation into the genomic barcode format, are proposed. These data with information about used primers that can be unique for genomes. A comparative analysis of genomic barcodes of species of related plant species was carried out in order to classify them on the level of species and lines in the future. A pairwise analysis of the location of the same or similar amplicons within different subgenomes and genomes is presented. The genomes of wheat and Aegilops in FASTA files format are presented as the research samples. The proposed method makes possible to predict the success of the multiplex polymerase chain reaction using special primers in the laboratory. This technology allows the analysis of the entire genomic DNA, rather than fragments of the genome.

Key words: *genomic barcoding, data digitization, multiplex PCR, RAPD-analysis, computer simulation, genome.*