

МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА РЕСПУБЛИКИ КАЗАХСТАН



Западно-Казахстанский аграрно-технический университет
имени Жангир хана

Бейшова И.С.
Гинятов Н.С.
Ульянов В.А.
Ульянова Т.В.
Ковальчук А.М.
Сариев Б.Т.
Сидарова А.Ж.

РЕКОМЕНДАЦИИ
ПО ОПРЕДЕЛЕНИЮ ГЕНЕТИЧЕСКОЙ СТРУКТУРЫ
ПОПУЛЯЦИЙ ОСЕТРОВЫХ ВИДОВ РЫБ И ИХ ГИБРИДОВ

Уральск, 2024

УДК 639.3
ББК 47.29-31
Р 36

**Рекомендовано к изданию НТС НАО «Западно-Казахстанский
аграрно-технический университет имени Жангир хана»
(протокол № 1 от «25» сентября 2024 г.)**

РЕЦЕНЗЕНТЫ:

Аубакирова Г.А. – НАО «Казахский агротехнический исследовательский университет имени С.Сейфуллина», заведующий кафедрой «Охотоведения и рыбного хозяйства», PhD, ассоциированный профессор

Туменов А.Н. – ТОО «Научно-производственный центр рыбного хозяйства», директор Западно-Казахстанского филиала, PhD

Бейшова И.С., Гинятов Н.С., Ульянов В.А. и др.

Р 36 Рекомендации по определению генетической структуры популяций осетровых видов рыб и их гибридов: Рекомендация / И.С. Бейшова, Н.С. Гинятов, В.А. Ульянов, Т.В. Ульянова, А.М. Ковальчук, Б.Т. Сариев, А.Ж. Сидарова – Уральск: НАО «Западно-Казахстанский аграрно-технический университет имени Жангир хана», 2024. – 51 с.

ISBN 978-601-319-528-5

Рекомендации основаны на исследованиях НИР по бюджетной программе 217 "Развитие науки", подпрограмме 102 "Грантовое финансирование научных исследований", конкурсу научных и (или) научно-технических проектов на 2022-2024 годы, проекту ИРН AP14870980 «Изучение специфических особенностей генетической структуры осетровых рыб и их гибридов, выращиваемых в установках замкнутого водоснабжения»

УДК 639.3
ББК 47.29-31

© Бейшова И.С., Гинятов Н.С., Ульянов В.А., Ульянова Т.В.,
Ковальчук А.М., Сариев Б.Т., Сидарова А.Ж., 2024
© НАО «Западно-Казахстанский аграрно-технический
университет имени Жангир хана», 2024

ISBN 978-601-319-528-5

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	4
1. Молекулярно-генетические маркеры.....	5
2. Отбор образцов и хранение биоматериала.....	10
3. Выделение ДНК.....	14
4. STR-генотипирование.....	15
5. Обработка результатов генотипирования.....	17
6. Популяционно-генетический анализ в Excel.....	22
7. Кластерный анализ с использованием программы STRUCTURE.....	25
8. Выбор оптимального значения K.....	28
8.1 Structure Harvester.....	28
8.1.1 Веб-интерфейс программы Structure Harvester.....	28
8.1.2 Портативная версия программы Structure Harvester.....	30
8.2 CLUMPAK.....	31
9. Объединение результатов кластеризации.....	33
9.1 CLUMPP.....	33
9.2 CLUMPAK.....	37
10. Построение филогенетических деревьев в программе SplitsTree4.....	38
ЗАКЛЮЧЕНИЕ.....	42
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....	44

ВВЕДЕНИЕ

По данным Международного агентства по охране окружающей среды, осетровые являются группой видов, находящихся под наибольшей угрозой исчезновения [1]. По этой причине осетровые являются объектами работ, направленных на сохранение видов, их часто включают программы по пополнению естественных популяций молодь, полученной в неволе; такая деятельность по сохранению *ex-situ* (стратегия, при которой ресурсы генофонда тех или иных видов содержатся в искусственных условиях (семена, пыльца, сперма, замороженные клетки, отдельные организмы в ботанических садах и зоопарках) обязательно должна сопровождаться исследованиями, направленными на сохранение генетического разнообразия посредством долгосрочных программ селекции [2]. Однако генетический анализ осетровых имеют ряд сложностей, связанных с особенностями их генома и различными уровнями пloidности из-за независимых событий полногеномной дубликации [3, 4].

Биоразнообразие в пресноводных экосистемах сократилось гораздо сильнее, чем в наземных и морских экосистемах [5-7]. Вследствие деятельности человека и расширения поселений в направлении рек многочисленные пресноводные рыбы оказались под угрозой исчезновения из-за разделения среды обитания (например, из-за плотин и дамб), загрязнения и чрезмерного вылова рыбы [8-10]. В частности, рыбы, обитающие в реках, чувствительны к воздействию физических, химических и биологических факторов, быстрые изменения условий окружающей среды повышают вероятность исчезновения видов [11-13]. По этой причине разрабатываются планы восстановления, чтобы предотвратить исчезновение редких видов [14-17].

Ранее при составлении подобных планов не учитывались генетические исследования, направленные на установление состояния аллельного разнообразия [16]. Однако необходимость подобных исследований была неоднократно доказана для эффективного восстановления исчезающих видов

[18, 19]. Соответственно, в последние десятилетия генетические технологии совершенствовались и применялись при планировании сохранения видов, но при этом сохраняется немало проблем, связанных как с высокой стоимостью подобных исследований, так и с недостаточностью образовательных материалов по данной теме [17, 20, 21].

Генетическое разнообразие является ключевым показателем, который позволяет видам адаптироваться к изменяющимся условиям окружающей среде и сохранять свой эволюционный потенциал [18, 22, 23]. Сокращение и фрагментация ареалов обитания ограничивает поток генов, тем самым увеличивая генетический инбридинг [18, 24]. Рост инбридинга вызывает инбредную депрессию, которая в свою очередь еще больше ускоряет сокращение и вымирание популяции [25-27]. Таким образом, в силу целого ряда причин эффективность естественного воспроизводства осетровых в р. Урал в настоящее время сведена практически к нулю.

Настоящие рекомендации разработаны НАО «Западно-Казахстанский аграрно-технический университет имени Жангир хана» в рамках выполнения проекта: АР14870980 «Изучение специфических особенностей генетической структуры осетровых рыб и их гибридов, выращиваемых в установках замкнутого водоснабжения» (№ гос.регистрации 0122РК00243).

1. Молекулярно-генетические маркеры

Генетические маркеры представляют собой участки ДНК, которые позволяют отслеживать наследование различных генетических характеристик, присутствующих на хромосоме, определенном сегменте хромосомы или конкретном местоположении гена в семейной линии. Генетические маркеры характеризуются двумя фундаментальными особенностями: наследуемость и идентифицируемость (можно легко найти интересующий маркер и четко отличить один маркер от другого). Это делает генетические маркеры жизненно важным инструментом в генетических исследованиях и анализе.

Молекулярно-генетические маркеры не изменяются под воздействием таких факторов, как стадия развития, время года или условия окружающей среды. Они многочисленны, распределены по всему геному, демонстрируют высокую степень полиморфизма и множество аллельных вариаций.

Многие маркеры оказывают нейтральное действие, не влияя на проявление целевых признаков и не имея связи с нежелательными характеристиками. Существует множество быстрых и относительно недорогих методов их обнаружения.

Типы молекулярных маркеров

Первое поколение - ПДРФ (полиморфизм длин рестрикционных фрагментов)

Метод ПДРФ позволяет обнаружить изменения в последовательности ДНК, возникающие в результате таких процессов, как замены оснований, вставки, делеции или дупликации, которые влияют на сайты узнавания эндонуклеаз рестрикции в геноме [28]. Применение маркеров первого поколения со временем сократилось.

Принцип исследования методом ПДРФ предполагает расщепление геномной ДНК разных особей с помощью эндонуклеаз рестрикции, в результате чего образуются фрагменты ДНК разного размера. Эти фрагменты разделяют при помощи гель-электрофореза и переносят на гибридизационную мембрану. Зондом служит специфический фрагмент ДНК, меченный изотопами или неизотопами. Зонд гибридизуется с фрагментами, расщепленными ферментом, выявляя различия в длине фрагментов с гомологичными последовательностями зонду [29].

Преимуществами ПДРФ маркеров являются их кодоминантный характер, высокая воспроизводимость и стабильность.

Однако ПДРФ маркеры имеют и ряд недостатков, включая сложные и продолжительные экспериментальные процедуры, высокую стоимость и ограниченную пригодность для крупномасштабных проектов по генотипированию. Кроме того, использование радиоактивных изотопов для

детекции фрагментов ДНК является небезопасным для исследовательского персонала и приводит к различного рода загрязнениям. Нерадиоактивные методы мечения, как правило, более дорогие, имеют более слабые сигналы гибридизации и сниженную чувствительность.

Второе поколение – молекулярные маркеры на основе ПЦР (полимеразно-цепная реакция)

Технологии молекулярных маркеров второго поколения в настоящее время наиболее широко используются в молекулярной генетике. В эту категорию входят маркеры, обнаруживаемые при помощи прямой ПЦР, так и ПЦР в сочетании с подходами ферментативного расщепления.

RAPD (Random Amplified Polymorphic DNA) - случайно амплифицированная полиморфная ДНК, предполагает использование коротких случайных праймеров, обычно длиной 8-10 оснований, для амплификации геномной ДНК. Когда расстояние между сайтами связывания двух праймеров соответствует условиям ПЦР-амплификации, образуется продукт амплификации. Любые вставки, делеции или мутации, происходящие между сайтами отжига праймера или на сайте его связывания, приводят к присутствию или отсутствию полос при электрофоретическом разделении продуктов ПЦР, что и приводит к появлению маркеров RAPD [30, 31].

Преимуществами RAPD являются техническая простота и отсутствие необходимости предварительного секвенирования последовательности ДНК, один и тот же набор праймеров можно применять для геномного анализа различных организмов [32, 33]. Недостатком RAPD маркеров является тот факт, что полиморфизмы обнаруживаются только как наличие или отсутствие полосы определенной молекулярной массы, без информации о гетерозиготности, а также наличие проблем с воспроизводимостью экспериментальных данных [34].

SSR маркеры (Simple Sequence Repeats – простые повторяющиеся последовательности), также известные как микросателлиты или короткие tandemные повторы (STR - short tandem repeats), представляют собой класс

последовательностей ДНК, состоящих из повторяющегося набора нуклеотидов, как правило 1-6 единиц. Эти повторяющиеся последовательности широко распространены в геномах эукариот и некоторых прокариот, также обнаруживаются в ядерной ДНК, ДНК хлоропластов и митохондриальной ДНК.

Принцип SSR маркеров заключается в высокой консервативности однокопийных последовательностей на обеих сторонах микросателлитной ДНК. Таким образом, на основе этих фланкирующих последовательностей можно создать праймеры для амплификации ПЦР, а размер продуктов ПЦР можно определить с помощью электрофореза. Высокий полиморфизм маркеров SSR в первую очередь обусловлен вариациями числа tandemных повторов внутри основной последовательности, таким образом, они широко используются в исследованиях по изучению генетического разнообразия, генетической структуры, эффектов «бутылочного горлышка» и генетического потока среди популяций [35-43].

Обычно микросателлитные локусы имеют 10 и более аллельных вариантов и гетерозиготность выше 60%, даже при небольшой выборке. Эти качества, в совокупности с широкой распространённостью микросателлитов по всему геному, делают их наиболее привлекательными генетическими маркерами для высокоэффективного генетического картирования [44, 45].

Преимущества SSR маркеров: представляют собой кодоминантные маркеры, позволяющие дифференцировать гетерозиготы и гомозиготы, требуют небольшого количества ДНК-материала и не слишком чувствительны к качеству ДНК, демонстрируют отличную воспроизводимость, высокую степень аллельного разнообразия (полиморфизма).

Ограничения SSR маркеров: знание последовательности ДНК на обоих концах повторяющегося мотива имеет важное значение. В тех случаях, когда эта информация недоступна в базах данных ДНК, становится необходимым секвенирование и разработка праймеров, что приводит к увеличению затрат на

разработку. Не очень высокая степень информативности анализа – за одну реакцию можно проамплифицировать около 20 STR маркеров.

Третье поколение - SNP-маркеры

SNP (Single Nucleotide Polymorphism - однонуклеотидный полиморфизм) отличия последовательности ДНК размером в один нуклеотид (А, Т, G или С). Оно возникает в результате таких изменений, как переходы, трансверсии, вставки, делеции и т. д. [46-49].

SNP можно обнаружить и проанализировать с помощью различных методов. Когда известна последовательность ДНК, ее можно сравнить с базами данных, содержащих эталонные геномы или геномы близкородственных видов, для обнаружения SNP. Кроме того, если SNP расположены в определенных сайтах узнавания ферментов рестрикции, на основе последовательностей, фланкирующих SNP, можно создать праймеры для ПЦР-амплификации. Продукты ПЦР можно расщепить ферментами рестрикции и подвергнуть электрофорезу для превращения их в кодоминантные маркеры CAPS. Альтернативно, можно сконструировать праймеры для размещения SNP на 3'-конце, а наличие или отсутствие амплифицированных полос используется для преобразования их в ПЦР-маркеры. Более современные методы включают секвенирование ДНК или использование методов гибридизации ДНК-чипов для обнаружения [50, 51].

Преимущества: SNP-маркеры многочисленны и широко распространены в геномах, обладают высокой стабильностью, кодоминантным характером и подходят для быстрого и масштабного скрининга.

Недостатки: высокая стоимость при использовании методов секвенирования или гибридизации с помощью ДНК-чипов, низкий уровень получаемой информации по сравнению с высокополиморфным микросателлитом, но эта проблема решается при использовании большого количества маркеров.

2. Отбор образцов и хранение биоматериала

Для оценки генетической структуры видов и популяций необходим надежный источник биологического материала. Источником биологического материала, а, следовательно, и ДНК могут служить волосы, фекалии, моча, сброшенные перья, кожа, чешуя, яичная скорлупа и др. Такой способ отбора проб кажется очень привлекательным, поскольку он позволяет получить биоматериал и провести генетический анализ особей без необходимости их поиска и вылова, проведения каких-либо манипуляций и наблюдений [52-54]. Однако данная стратегия обычно позволяет получить небольшое количество и низкое качество ДНК, содержит ДНК других организмов и не обеспечивает идентификацию образцов, что ограничивает потенциал этого метода.

Прижизненный отбор проб, подразумевающий взятие тканей (крови, кожи, чешуи, мышц) без критического повреждения животных, может использовать весь потенциал геномного анализа ДНК, включающий определение пола, оценку родства, определение эффективного размера популяции, уровня генетического полиморфизма и многое другое [55].

При исследовании геномов рыб, наиболее часто, для получения большого количества качественной ДНК используют печеночную и мышечную ткани, однако когда приходится иметь дело с популяциями или видами находящимися под угрозой исчезновения, данные способы получения биологического материала считаются неприменимыми. В этом случае для выделения ДНК используют фрагменты плавников или чешую рыб. Данный метод отбора проб ДНК обеспечивает исследователя большим количеством высококачественной ДНК, которую можно использовать в качестве матрицы для полимеразной цепной реакции (ПЦР), и позволяет изучать особи без нанесения вреда их здоровью.

В нашем исследовании для взятия биоматериала выбран способ прижизненного отбора биологических образцов, без использования методики декапитации осетров.

В качестве исходного материала для проведения генетических исследований по определению происхождения рыб взяты эктодермальная ткань – фрагменты грудного плавника осетров (рисунок 1).



Рисунок 1 – Отбор фрагмента лучевой части грудного плавника

Участок плавника механически очищался от естественной слизи и других загрязнений ватно-марлевым тампоном, смоченном в 70%-м спирте. Далее стерильными инструментами производится отбор биологического материала, не допуская использования скальпелей и пинцетов, загрязненных остатками тканей от предыдущих образцов. От каждой особи было взято две пробы, одна из них была зафиксирована в 96% этиловом спирте на месте забора материала, другая помещена в пластиковую пробирку без фиксации спиртом. К каждому взятому экспертному материалу, направляемых для выделения и последующей очистки ДНК, присвоен идентификационный номер, с целью

недопущения смешения проб от разных особей. В обязательном порядке проводится сканирование пойманных особей на наличие чипа и чипирование в случае его отсутствия, номера которых используются для дальнейшей идентификации и паспортизации (рисунок 2). На месте составляется и подписывается акт отбора биоматериала (рисунок 3).



Рисунок 2 – Установка подкожных чипов для идентификации особей

АКТ
отбора биологического материала и чипирования осетровых рыб

Отбор биологического материала был произведен в рамках реализации научного проекта АР14870980 «Изучение специфических особенностей генетической структуры осетровых рыб и их гибридов, выращиваемых в установках замкнутого водоснабжения» руководителем проекта Н.С. Гинаятовым и исполнителями В.А. Ульяновым и А.М. Ковальчук.

Чипирование осетровых рыб было проведено исполнителем проекта Б.Т. Сариевым в присутствии техника-рыбовода лаборатории аквакультуры М. Энгельса.

Отобраны образцы биологического материала для микросателлитного анализа согласно инструкции по сбору образцов тканей осетровых рыб для молекулярно-генетического анализа

Место отбора: лаборатория аквакультуры НАО «Западно-Казахстанского аграрно-технического университета имени Жангир хана»

Дата и время отбора: 20 сентября 2022 г, 14:40 ч.

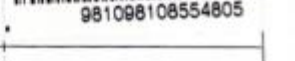
№ п/п	Вид рыбы	№ образца	Вид материала	Ед. изм.	Кол-во	№ микрочипа
1.	Русский осетр	1	Фрагмент плавника	шт	1	 981098108554567
2.	Русский осетр	2	Фрагмент плавника	шт	1	 Sterilization expiry date: 03/2026 981098108554577
3.	Русский осетр	3	Фрагмент плавника	шт	1	 Sterilization expiry date: 03/2026 981098108554542
4.	Русский осетр	4	Фрагмент плавника	шт	1	 981098108554549
5.	Русский осетр	5	Фрагмент плавника	шт	1	 981098108554533
6.	Русский осетр	6	Фрагмент плавника	шт	1	 981098108554415
7.	Русский осетр	7	Фрагмент плавника	шт	1	 981098108554451
8.	Русский осетр	8	Фрагмент плавника	шт	1	 981098108554652
9.	Русский осетр	9	Фрагмент плавника	шт	1	 Sterilization expiry date: 03/2026 981098108554740
10.	Русский осетр	10	Фрагмент плавника	шт	1	 981098108554805
11.	Русский осетр	11	Фрагмент плавника	шт	1	 981098108554657
12.	Русский осетр	12	Фрагмент плавника	шт	1	 981098108554600

Рисунок 3 – Форма акта отбора биоматериала

Все отобранные пробы необходимо как можно быстрее поместить в низкотемпературный холодильник для предотвращения порчи и длительного хранения.

3. Выделение ДНК

В нашем исследовании ДНК была экстрагирована из тканей плавников с использованием коммерческого набора «ДНК-Экстран-2» (Компания ООО «СИНТОЛ», РФ). Так же подойдут любые методы экстракции ДНК и коммерческие наборы, предназначенные для выделения нуклеиновых кислот из тканей животного происхождения.

Коммерческий набор «ДНК-Экстран-2» использовался согласно инструкции изготовителя. Данный набор предназначен для выделения геномной ДНК с высоким выходом и чистотой из самых разнообразных типов образцов (рисунок 4).



Рисунок 4 - Набор для выделения ДНК «ДНК-Экстран-2»

Работа проводится в зоне выделения нуклеиновых кислот.

Для работы с данным коммерческим набором требуется наличие минимально необходимого оборудования (термостат, микроцентрифуга; вортекс; дозаторы на 1000, 100, 10 мкл; штатив для пробирок) и расходных материалов (одноразовые наконечники (на 1000, 100, 10 мкл), пластиковые пробирки 1,5 мл, перчатки резиновые или латексные, неопудренные). Для

снижения риска кросс-контаминации все манипуляции производят в перчатках.

Полученный раствор ДНК требуется хранить при -20°C . Допустимо кратковременное хранение раствора ДНК при $+4^{\circ}\text{C}$.

Измерение концентрации ДНК. Измерение концентрации ДНК проводили на спектрофотометре Agilent Cary 60 (рисунок 5).



Рисунок 5 - Спектрофотометр Agilent Cary 60

Показателем чистоты служило значение соотношения коэффициента экстинкции при длине волны 260 нм (D260) к значению при длине волны 280 нм (D280). Так же допустимо пользоваться любым аналогичным оборудованием, предназначенным для измерения концентрации нуклеиновых кислот.

4. STR-генотипирование

Для генотипирования нами был использован набор реагентов «ГенЭксперт-Осетр» (Компания ООО «СИНТОЛ», РФ) предназначенный для генетической паспортизации и определения родства осетровых видов рыб.

Генотипирование проводилось по 7 микросателлитным локусам (AoxD161, AfuG41, LS19, AfuG135, AfuG37, Spl173 и AoxD165), характеристики которых указаны в таблице 1. Нами были выбраны именно эти локусы, так как они амплифицируются у большинства особей осетровых видов рыб, результаты генотипирования по данным локусам можно воспроизвести в условиях других лабораторий, все перечисленные локусы амплифицируются в мультиплексной ПЦР.

Применяемые сегодня мультиплексные STR наборы используют преимущества нескольких флуоресцентных красителей, которые позволяют амплифицировать гораздо большее количество локусов в одной реакции.

Таблица 1 – Характеристика микросателлитных локусов

Локус	Последовательность праймеров (5'-3')	Длина ампликонов, п.н.	Структура повтора	Ссылка на источник
AoxD161	F: GTTTGAAATGATTGAGAAAATGC R: TGAGACAGACACTCTAGTTAAACAGC	90-153	ATCT	[56]
Afug41	F: TGACGCACAGTAGTATTATTTATG R: TGATGTTTGCTGAGGCTTTTC	182-274	ATCT	[57]
LS19	F: CATCTTAGCCGTCTGGGTAC R: CAGGTCCCTAATACAATGGC	124-166	GTT	[58]
Afug135	F: GCCAATTTCCTGAAATATACCAG R: CGAAACCGCTTCAGACCTT	184-276	ATCT	[57]
Afug37	F: CAGGGAATCATGAGCACACG R: TGGCGCAGGATTTTGACAC	140-222	ATCT	[57]
Spl173	F: GGCTTTTGTCTGAAACGTCC R: TGGTGTGTGATTTTGAAGGC	226-324	ATCT	[59]
AoxD165	F: TTTGACAGCTCCTAAGTGATACC R: AAGCCCTACAACAAATGTCAC	158-244	ATCT	[56]

Реакцию амплификации проводили согласно инструкции производителя набора для генотипирования на термоциклере ProFlex PCR system (Applied Biosystems, США) (рисунок 6).

Разделение полученных фрагментов ДНК проводилось методом капиллярного электрофореза в полимерном геле с использованием генетического анализатора AB 3500 (Applied Biosystems, США) (рисунок 6).



Рисунок 6 – Трехблочный амплификатор ProFlex PCR system и генетический анализатор 3500 Applied Biosystems

Для определения генетической структуры какой-либо группы осетровых можно использовать любую подобранную панель микросателлитных маркеров и набор оборудования, позволяющий провести генетическое типирование и считать полученные результаты.

5. Обработка результатов генотипирования

Интерпретацию результатов анализа проводили при помощи программного обеспечения GeneMapper 6.0 (Applied Biosystems).

Данные, полученные с помощью анализатора 3500 Applied Biosystems или других аналогичных приборов, обычно представлены в виде пиков на электрофореграмме, которые соответствуют аллелям различным STR локусов, амплифицированных из образца ДНК. Эти пики представляют собой зависимость интенсивности флуоресцентного сигнала от времени прохождения детектора в системе капиллярного электрофореза. Компьютерные программы, подобные GeneMapper, позволяют преобразовать

эти флуоресцентные сигналы в «аллельные пики» с числовыми характеристиками (рисунок 7).

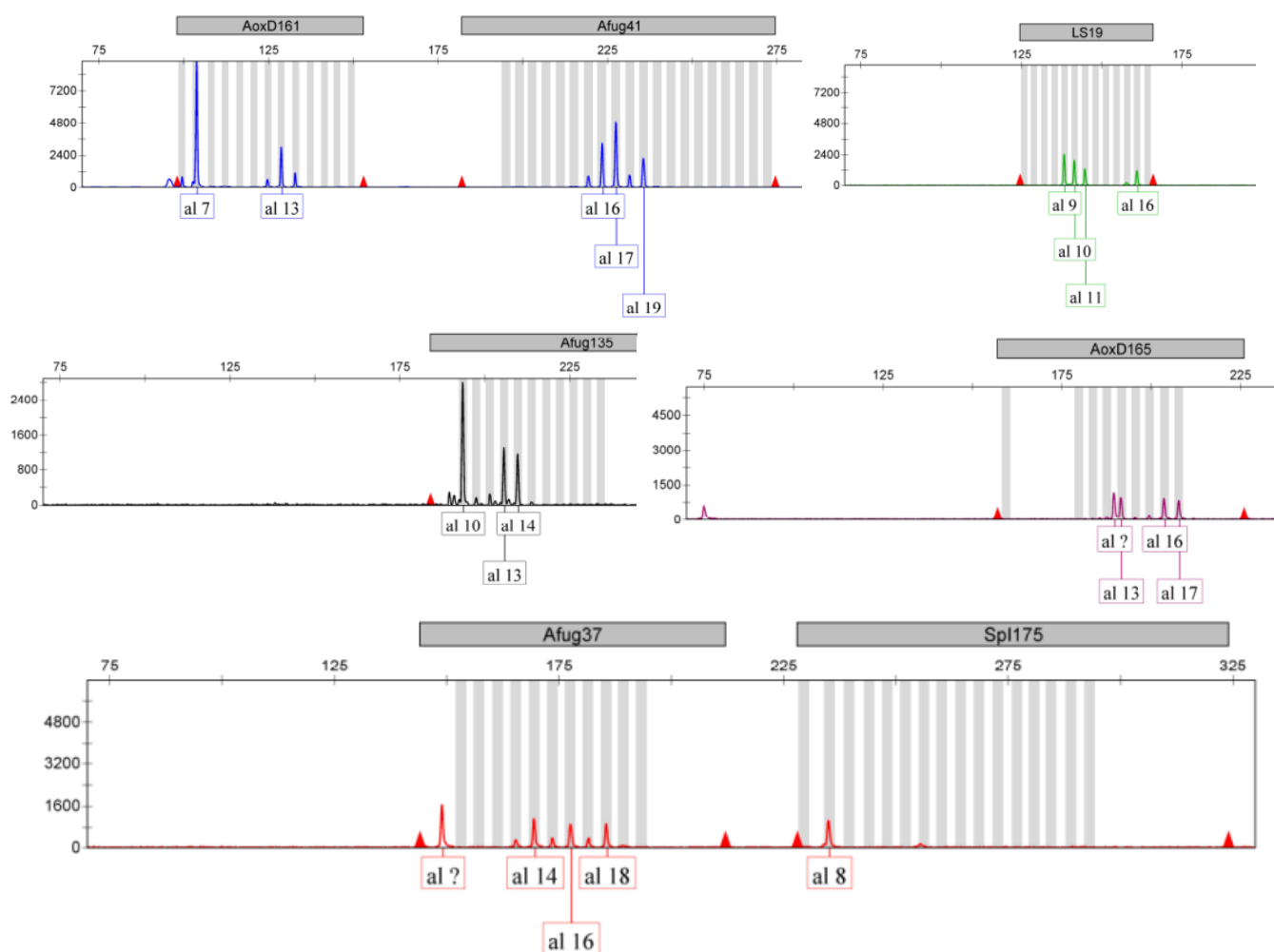


Рисунок 7 – Электрофореграмма результатов STR-генотипирования образца № 1 (№ микрочипа 643110800352541)

Как правило одновременно с образцом в лунку помещается размерный стандарт (в нашем случае СД-450) меченный красителем другого цвета, благодаря которому производится расчет длины детектируемого фрагмента ДНК. Размерный стандарт СД-450 содержит 22 фрагмента ДНК размером от 60 до 450 пар оснований и помеченных флуоресцентным красителем SY660 (спектральным аналогом красителя LIZ).

Производитель наборов для генотипирования предоставляет пользователям информацию о номенклатуре аллелей, как правило, каждая

аллель отражает количество повторов в данном локусе. Программное обеспечение позволяет преобразовывать все аллельные пики в обрабатываемых образцах из размера ДНК (относительно общего внутреннего размерного стандарта) в число повторов (рисунок 7). Обычно, по сторонам от каждой аллели, создается диапазон размером 0,5 п.н., называемый «бин» (bin), этот диапазон позволяет дифференцировать продукты ПЦР, отличающиеся друг от друга на 1 нуклеотид и, следовательно, отличать разные аллели [60].

Любой пик с одинаковой цветовой меткой, размер которого находится в пределах бина, обозначается программой генотипирования как данный аллель. Аналитик, выполняя анализ профиля ДНК, оценивает, действительно ли отмеченный пик является аллелем или ошибкой, например, продуктом «заикания» или усилением сигнала от другого красителя.

Электрофореграммы могут содержать и дополнительные пики, которые будут отличаться от истинных пиков. Такие пики могут возникать по ряду причин, связанных со структурными особенностями микросателлитных повторов, используемыми наборами реагентов (особенности праймеров, полимеразы и т.д.), расходными материалами для генетического анализатора, изменяющимися условиями окружающей среды (температура, влажность) и технологией обнаружения флуоресцентных сигналов. Специалисту, выполняющему интерпретацию результатов необходимо уметь выявлять и исключать такие пики.

Ошибочные пики, не являющиеся аллелями, подразделяются на несколько видов:

«Statter» - наиболее распространенный вид дополнительных пиков на электрофореграмме, возникает вследствие ошибок амплификации и обычно короче основного пика на 2 или 4 нуклеотида (рисунок 8). Как правило статтеры не превышают 10% от основного пика, однако если аллельный пик выходит за пределы диапазона интерпретации (при высокой концентрации ДНК-матрицы) высота статтера отображается больше, чем она есть на самом деле.

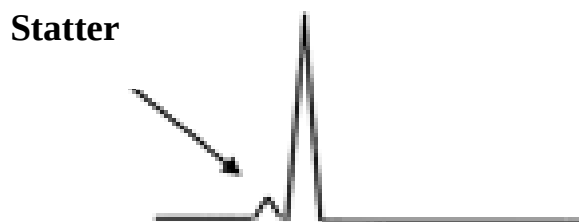


Рисунок 8 – Схематическая иллюстрация «статтера»

Неполное достраивание фрагмента ДНК – возникает в случае избытка ДНК-матрицы в полимеразной цепной реакции или при недостаточной оптимизации условий амплифицирования, как правило, данная ошибка возникает при присоединении аденина (A) (рисунок 9).

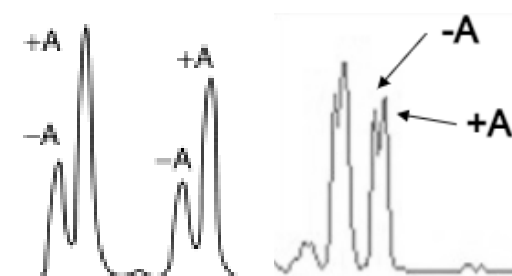


Рисунок 9 – Варианты пиков при неполном достраивании аденина (A)

Ошибка матрицы – является результатом неспособности детектора четко различать цвета красителей, используемых для маркировки ампликонов. Проявляется в спектральном перекрытии, пик одного цвета «подтягивается» или «просачивается» в другие цвета в результате выхода за пределы обнаружения (перегрузка ДНК-матрицы) (рисунок 10).

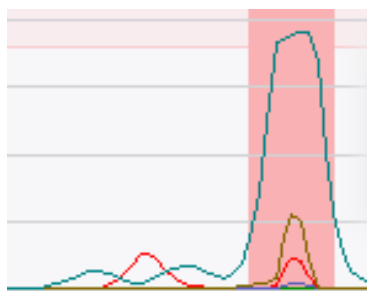


Рисунок 10 – Пример «подтягивания» цветов

Капли красителя – возникают в случае, когда флуоресцентные красители отрываются от соответствующих праймеров и самостоятельно мигрируют через капилляр во время электрофореза. Большинство таких красителей удаляются во время очистки праймеров, однако если по каким-либо причинам нарушается технологическая цепочка производства наборов, это приводит к возникновению таких ошибок. Пики образуются довольно широкие, имеют спектр одного из красителей и легко отличимы от других пиков (рисунок 11).



Рисунок 11 – Капля красителя на электрофореграмме

Пузырьки воздуха, кристаллы мочевины или скачки напряжения также могут привести к появлению ложных пиков на генетическом анализаторе, называемыми – spike (рисунок 12). Такие пики обычно резкие, тонкие и обладают одинаковой интенсивностью для всех цветов, как правило они исчезают если повторно переставить образец.

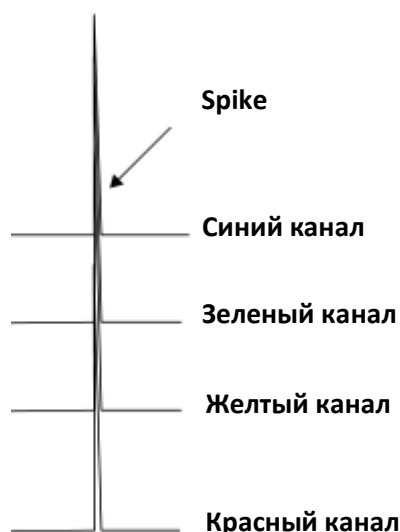


Рисунок 12 – Схематическое изображение помех, возникающих из-за пузырьков воздуха, кристаллов мочевины или скачков напряжения

Смешивание образцов или контаминация – представляют собой значительную проблему при интерпретации результатов генотипирования. При смешивании двух или более образцов на электрофореграмме во всех или некоторых локусах присутствуют посторонние пики, они могут быть меньшей интенсивности чем основной образец или аналогичны ему, в случае контаминации определенные пики появляются в одном месте во всей партии поставленных образцов. Электрофореграммы с признаками смешивания и контаминации не подлежат анализу и должны быть анализированы повторно, до устранения указанных дефектов.

Полученные числовые данные после интерпретации можно экспортировать из программы GeneMapper в формате txt, которые в дальнейшем преобразуются в табличный вариант с помощью различных приложений, например Microsoft Excel, для дальнейшего анализа.

6. Популяционно-генетический анализ в Excel

Существует большое количество программ, позволяющих проводить популяционно-генетический анализ, большинство из которых можно бесплатно загрузить с веб-сайтов их разработчиков. Некоторые из них работают на платформах подобных Java (STRUCTURE, Arlequin, Genepop и др.), предоставляя удобный пользовательский интерфейс, другие же представляют собой макросы или надстройки в Microsoft Excel (GenAlEx и Microsatellite Toolkit) [61].

Часть из таких программ предназначена для комплексного статистического анализа, однако большинство создается для расчета конкретного параметра, поэтому выбор используемых программ зависит от задач проекта и объекта исследования.

GenAlEx (Genetic Analysis in Excel) – широко используемая надстройка к Microsoft Excel для анализа генетико-популяционных данных. Данная надстройка позволяет проводить частотный анализ аллелей и генотипов,

анализ генетических дистанции (AMOVA, PCoA и др.), рассчитывать показатели структуры популяции и много другое. Очень важной характеристикой данного пакета является возможность экспорта данных в другие форматы (более 30), что позволяет быстро и эффективно проводить следующие этапы анализа, направленные на реализацию планов проекта.

Создание правильных входных файлов для аналитических программ является важным этапом, от которого зависит правильность получения результатов. Входной файл создается в программе Microsoft Excel, рекомендуется создать дополнительную вкладку, в которой будет указана информация об анализируемых образцах: шифр популяции, название вида согласно международной классификации, количество особей, дата отбора, дата генотипирования, тип популяции, место отбора и другие данные, которые могут быть полезны при интерпретации. Во вторую вкладку нужно внести генотипы согласно форме, указанной на рисунке 13.

5	10	4	2	3	2	3					
			popA	popB	popC	popD					
		LOC1		LOC2		LOC3		LOC4		LOC5	
popA1	popA	155	157	212	218	253	253	196	196	225	231
popA2	popA	155	155	212	220	253	263	178	196	231	231
popB1	popB	155	155	212	212	263	263	196	196	225	231
popB2	popB	157	157	212	218	259	263	196	196	225	231
popB3	popB	155	157	220	220	253	253	178	196	225	225
popC1	popC	155	157	212	220	253	259	196	196	225	225
popC2	popC	155	159	220	220	259	263	178	196	225	231
popD1	popD	157	157	212	212	245	245	196	196	225	231
popD2	popD	157	157	212	220	253	259	196	196	225	225
popD3	popD	155	157	212	212	245	253	196	196	225	225

Рисунок 13 – Формат входного файла для программы GenAlEx

В ячейке A1 указывается количество локусов, в ячейке B1 – общее количество образцов, C1 – количество популяций, D1 – количество особей в 1й популяции, E1 – количество особей во 2й популяции и т.д. Во второй строке, под количество особей в каждой популяции указывается шифр данной группы образцов, в третьей строке начиная с третьего столбца указываются

использованные для генотипирования локусы. Начиная с четвертой строки построчно указывается шифр образца, принадлежность к популяции и выявленный генотип особи, в случае, если данные для какого-либо локуса отсутствуют, то проставляются «0».

Загрузить программу GenAlEx можно с официального сайта разработчиков – <https://biology-assets.anu.edu.au/GenAlEx/Welcome.html>. На данном сайте также доступна подробная пользовательская инструкция с указанием всех используемых формул для каждого из расчетов. После загрузки и установки данной надстройки в Microsoft Excel появляется соответствующая вкладка (рисунок 14).

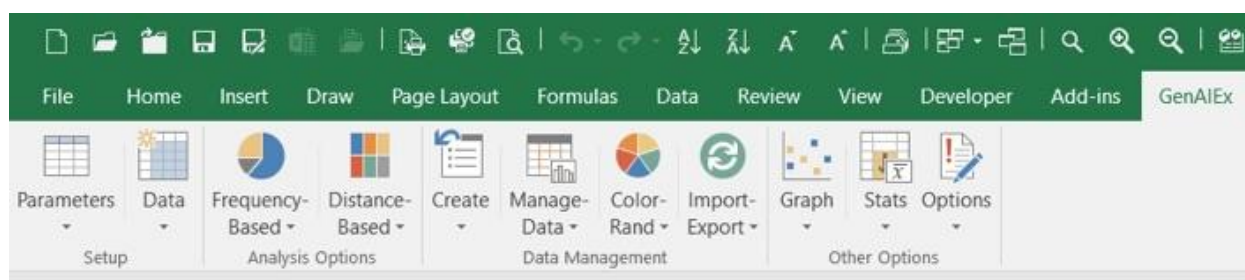


Рисунок 14 – Меню надстройки GenAlEx в Microsoft Excel

Необходимо открыть вкладку «Frequency-Based» – «Frequency» (либо выбрать другой тип анализа, в зависимости от поставленных задач), в открывшемся окне, проверить указанные данные и выбрать формат данных (бинарные данные, гаплоидные, диплоидные, кодоминирование). В зависимости от выбранного формата данных открывается соответствующее окно с возможными вариантами частотного анализа (Frequency Options), здесь нужно отметить необходимые значения. Полученные результаты будут отображаться на новых листах в Microsoft Excel.

Таким образом, можно быстро и точно рассчитать следующие показатели: частота аллелей, количество аллелей на локус (N_a), эффективное число аллелей (N_e), наблюдаемая (H_o) и ожидаемая (H_e) гетерозиготность, уровень фиксации Райта (F), коэффициенты инбридинга на разных уровнях

популяционной иерархии (Fis, Fit, Fst), показатель интенсивности потока генов (Nm), индекс генетической дистанции по Нею и многое другое. Данные будут представлены как в табличном, так и графическом виде.

7. Кластерный анализ с использованием программы STRUCTURE

Кластерный анализ – многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы [62].

Существует достаточно программ для проведения кластерного анализа генотипических данных, самой популярной из которых является STRUCTURE (https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/html/structure.html) [64]. В ней реализован байесовский алгоритм кластеризации генотипов в K кластеров с учетом априорной информации о географическом положении рассматриваемых популяций. Программа позволяет выделять группы, определять принадлежность особи к определенной группе, выявлять гибридизацию и промежуточные состояния.

Для анализа могут использоваться различные наборы данных (SNP, STR, RFLP, SSR и многие другие), включая фенотипические. Файл для импорта должен быть в формате .txt с разделением между числами с помощью табуляции и выглядеть следующим образом (рисунок 15).

		LOC1	LOC2	LOC3	LOC4	LOC5					
popA1	1	155	157	212	218	253	253	196	196	225	231
popA2	1	155	155	212	220	253	263	178	196	231	231
popB1	2	155	155	212	212	263	263	196	196	225	231
popB2	2	157	157	212	218	259	263	196	196	225	231
popB3	2	155	157	220	220	253	253	178	196	225	225
popC1	3	155	157	212	220	253	259	196	196	225	225
popC2	3	155	159	220	220	259	263	178	196	225	231
popD1	4	157	157	212	212	245	245	196	196	225	231
popD2	4	157	157	212	220	253	259	196	196	225	225
popD3	4	155	157	212	212	245	253	196	196	225	225

Рисунок 15 – Формат входного файла для программы STRUCTURE

Каждая особь располагается в отдельной строке, где в столбцах указываются локусы или другие цифровые характеристики объекта. В случае с диплоидными данными локусы будут расположены в двух столбцах. Не допускается использование дробных чисел.

В нашем случае, при работе с осетровыми, встречаются в основном тетраплоидные локусы, но также есть диплоидные, гекса- и октаплоидные локусы, рекомендуется использовать бинарные матрицы данных. В таких матрицах в первую строку вносятся все возможные значения для каждого локуса, а затем построчно заполняются показатели присутствия или отсутствия (1/0) этого значения для каждого из образцов (рисунок 16).

	AoxD161-90	AoxD161-100	AoxD161-104	AoxD161-108	AoxD161-110	AoxD161-112	AoxD161-116	AoxD161-120	AoxD161-124	AoxD161-128
Sample_01	0	0	1	0	0	0	0	0	0	1
Sample_02	0	0	1	0	0	1	0	0	0	1
Sample_03	0	0	1	0	0	1	0	0	0	1
Sample_04	0	0	1	0	0	0	0	0	1	0
Sample_05	0	0	0	0	0	1	1	0	0	1
Sample_06	0	0	1	0	0	1	0	0	0	1
Sample_07	0	0	1	0	0	1	0	0	0	0
Sample_08	0	0	0	0	0	1	0	0	1	0
Sample_09	0	0	0	0	0	1	0	0	1	1
Sample_10	0	0	0	0	0	1	0	0	1	0
Sample_11	0	0	0	0	0	1	1	0	0	1
Sample_12	0	0	0	0	0	1	0	0	1	1
Sample_13	0	0	1	0	0	1	0	0	1	0
Sample_14	0	0	0	0	0	0	0	0	1	0
Sample_15	0	0	0	0	0	0	0	0	1	1
Sample_16	0	0	1	0	0	1	0	0	1	0
Sample_17	0	0	1	0	0	1	0	0	0	1
Sample_18	0	0	1	0	0	1	0	0	0	0
Sample_19	0	0	1	0	0	1	0	0	1	0
Sample_20	0	0	1	0	0	1	0	0	0	1

Рисунок 16 – Бинарная матрица данных

Процесс создания проекта в STRUCTURE состоит из четырёх последовательных шагов, где нужно указать информацию об названии проекта, месте нахождения, дата файла, уровень плоидности, количество особей, локусов, наличие строки с названиями локусов, обозначение образцов, популяций и другие параметры.

После успешного создания проекта отобразится окно с исходными данными. Необходимо выбрать вкладку «Parameter Set» → «New», в открывшемся окне задаем необходимые нам параметры: Length of Burhin

Period (число повторов) – 100 000, Number of MCMC Reps after Burnin (число прожигов) – 1 000 000, ставим галочку в ячейке Allele Frequencies Independent, если аллели не связаны и задаем имя для сохранения выбранных параметров (выбор параметров зависит от задач и выбирается на усмотрение пользователя).

Далее выбираем вкладку «Project» → «Start a Job», открывается окно, в котором необходимо указать параметры запуска: значение «K» (на какое количество кластеров вы хотите разбить ваши данные, от и до) и количество итераций (рекомендуется 10), нажимаем кнопку «Start». В зависимости от выбранных параметров и мощности используемого компьютера процесс может занять от нескольких часов до суток.

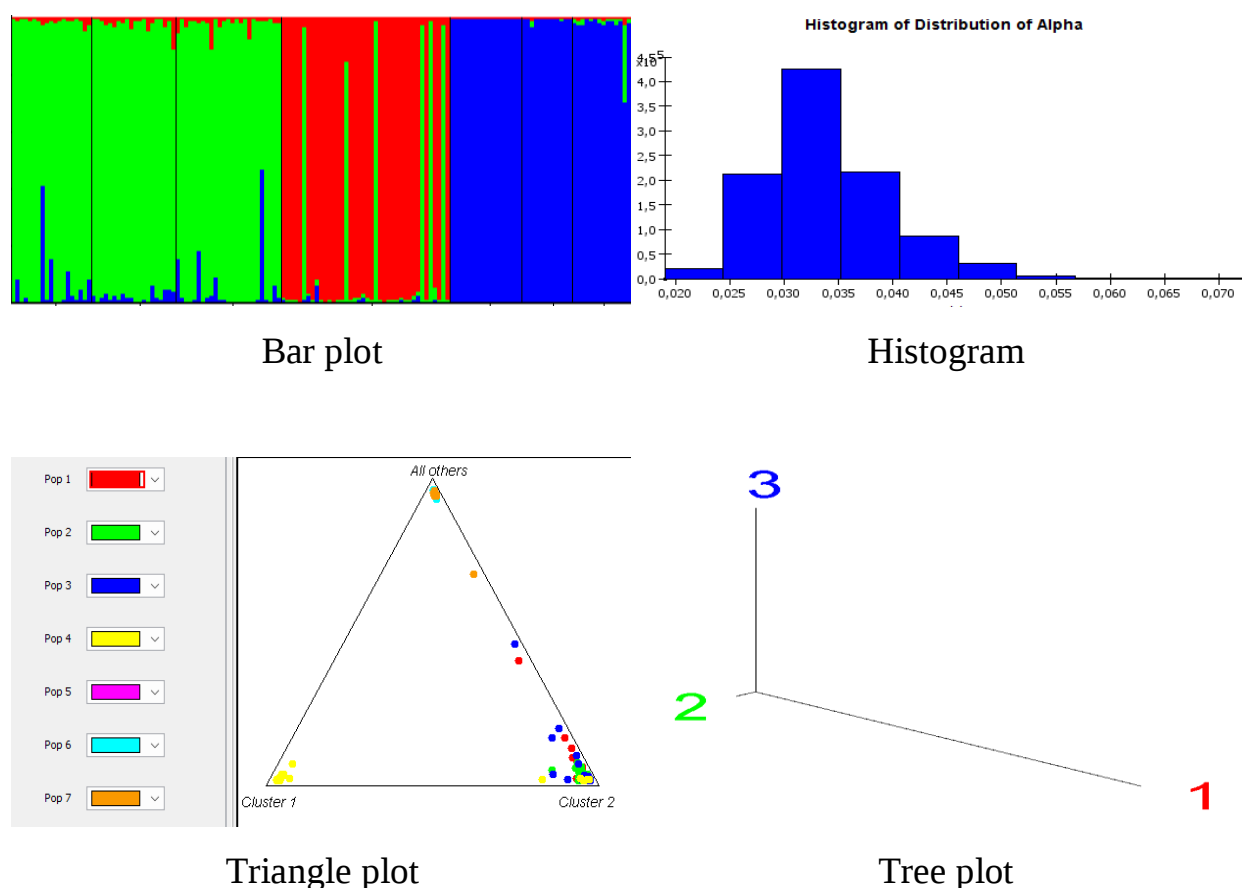


Рисунок 17 – Графическое отображение полученных данных в программе STRUCTURE

После окончания работы появится сообщение «Job is Completed!» и в левой окне программы будут доступны результаты всех прогонов, нажав на которые можно ознакомиться с полученными числовыми данными и представить их в виде графических моделей (рисунок 17): Bar plot (столбцовая диаграмма, гистограмма), Histogram (гистограмма), Triangle plot (треугольная диаграмма), Tree plot (дерево).

В программе существуют различные настройки, позволяющие настроить нужные параметры изображения.

Реализованный в STRUCTURE байесовский алгоритм позволяет идентифицировать генетически однородные группы особей. Однако данный алгоритм не способен обнаруживать истинное количество кластеров (K) в выборке [64]. Для математического определения наиболее оптимального значения K применяются методы Evanno G. и соавт. с помощью этих методов рассчитывается показатель Delta K и выявляется наиболее достоверное количество кластеров путем их последовательного перебора.

8. Выбор оптимального значения K

Провести расчет значения Delta K можно двумя способами: используя формулы, описанные в работе Evanno G. et al. (2005) [64]; с помощью программ в которых реализован упомянутый выше алгоритм.

8.1 Structure Harvester

8.1.1 Веб-интерфейс программы Structure Harvester

Для использования онлайн версии программы Structure Harvester необходимо перейти по ссылке (<http://taylor0.biology.ucla.edu/>

[structureHarvester/](http://structureharvester.org/)) и загрузить результаты полученные в программе STRUCTURE (рисунок 18).

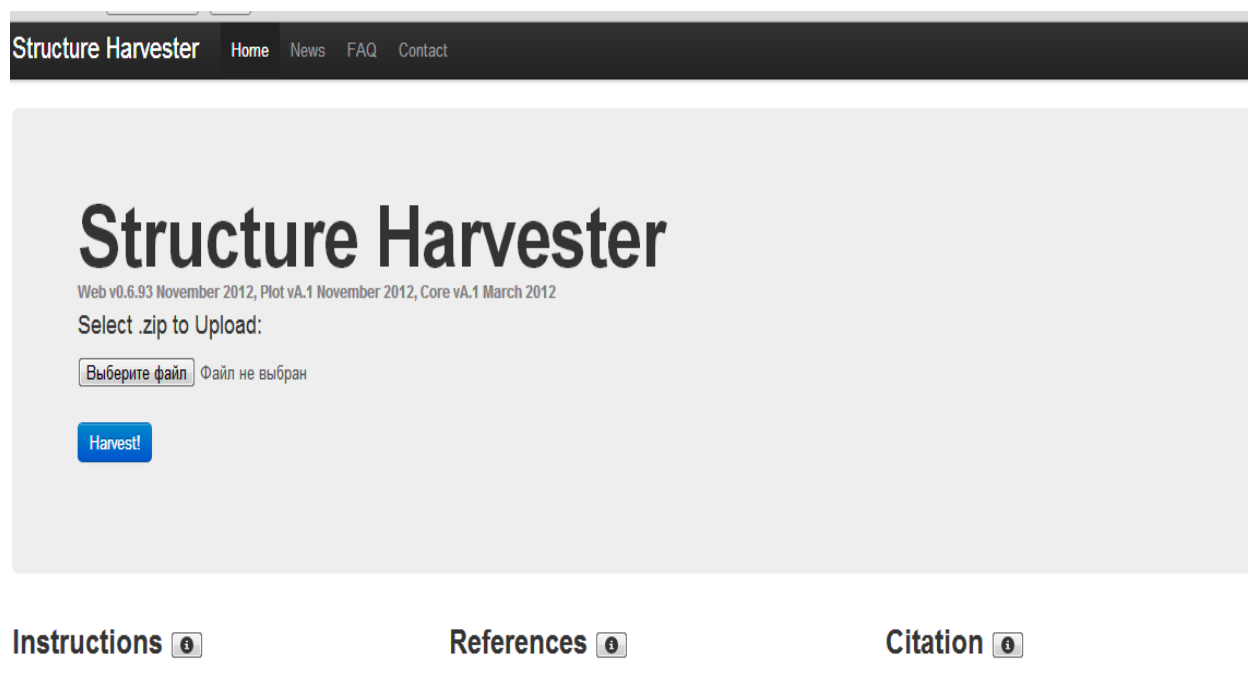
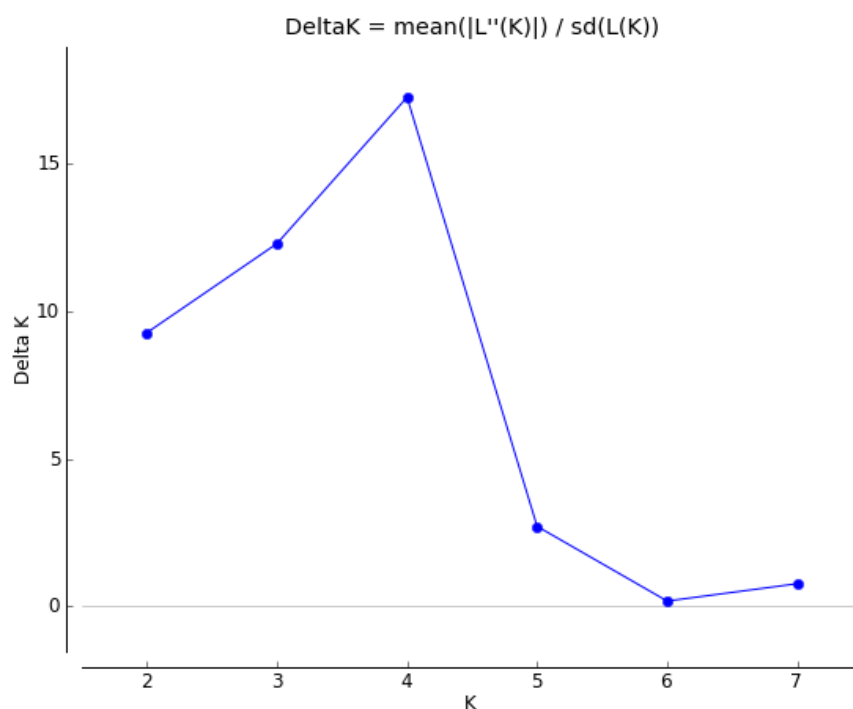


Рисунок 18 – Веб-интерфейс программы Structure Harvester

Результаты прогона в программе STRUCTURE находятся в указанной вами директории при создании проекта. В данной папке будут отображены файлы с выбранными параметрами и две дополнительные папки PlotData и Results. Папку Results добавляем в архив и загружаем в программу Structure Harvester, после завершения работы которого на экране отобразится серия графиков и таблиц, которые можно сохранить на свой компьютер и использовать в дальнейшем при интерпретации данных (рисунок 19).



K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	Ln''(K)	Delta K
1	10	-5077.960000	0.069921	—	—	—
2	10	-4539.440000	9.659216	538.520000	89.490000	9.264728
3	10	-4090.410000	24.444015	449.030000	300.200000	12.281125
4	10	-3941.580000	4.284027	148.830000	73.830000	17.233784
5	10	-3866.580000	43.486468	75.000000	117.770000	2.708199
6	10	-3909.350000	407.059218	-42.770000	73.150000	0.179704
7	10	-4025.270000	649.693651	-115.920000	498.460000	0.767223
8	10	-3642.730000	27.781011	382.540000	—	—

Рисунок 19 – Пример графика и табличных значений Delta K

Из представленного рисунка видно, что при работе программы STRUCTURE было выбрано от 1 до 8 кластеров, однако наиболее статистически достоверным является разделение выборки на 4 генетических кластера.

8.1.2 Портативная версия программы Structure Harvester

Портативная версия представляет собой скрипт на языке Python, который повторяет функциональность онлайн версии, но без возможности автоматического построения графиков.

Скачать можно перейдя по ссылке (<https://github.com/dentearl/structureHarvester/>), далее необходимо разархивировать и скопировать в новую папку файлы `structureHarvester.py` и `harvesterCore.py` (папку рекомендуется создать в той же директории, где расположен проект). В этой же папке создать еще две папки с именами «out» и «res», в папку `res` скопировать папку с результатами программы `STRUCTURE` и запустить терминал (имена папок можно менять на усмотрение пользователя, но при этом их также необходимо будет изменить в следующих командах на терминале).

Важно: для работы скрипта необходим установленный на компьютере Python3! Проверить установку Python на компьютере можно используя в терминале команду «python3», если язык установлен отобразится сообщение с версией, датой и временем установки.

Если все необходимые условия выполнены, на терминале нужно перейти в директорию где сохранены созданные папки и файлы `structureHarvester.py` `harvesterCore.py`, затем ввести команду для запуска программы (`./structureHarvester.py --dir=./res/Results --out=./out/ --evanno --clumpp`). После этого в папку `out` будут сохранены исходящие файлы с полученными расчетами. В файле «evanno» будет отражена следующая информация: когда был создан документ, какой программой, откуда была скачана программа, ссылка для цитирования метода расчетов и сами полученные результаты (Mean LnP(K), Stdev LnP(K), Ln'(K), |Ln''(K)| и Delta K) из которых можно построить необходимые графики. Помимо указанного файла в папке будут содержаться файлы (расширение `INDFILE` и `POPFIL`) с результатами перебора данных для каждого кластера по отдельности.

8.2 CLUMPAK

CLUMPAK - Clustering Markov Packager Across K – онлайн программа, разработанная с целью визуализации и обработки результатов, полученных в приложениях подобных `STRUCTURE`. Воспользоваться ей можно перейдя по

ссылке (<https://tau.evolveq.net/clumpak/index.html>). Для выполнения необходимых нам действий нужно перейти на вкладку Best K, загрузить результаты из программы STRUCTURE в zip архиве и нажать на кнопку Submit Form (рисунок 20). После этого программа в течении некоторого времени будет выполнять расчеты и выдаст результат в новом окне, которые можно будет сохранить на свой компьютер.

Main Pipeline Distruct for many K's Compare **Best K** Download Help Contact & Citing Issues

Prof. Itay Mayrose Lab - Plant Evolution, bioinformatics, & comparative genomics

CLUMPAK - CLUSTER MARKOV PACKAGER ACROSS K

This is a beta version.

We are working on improving CLUMPAK for enhanced flexibility and a better user interface. If you run into problems please email evolveq@tauex.tau.ac.il and specify the job ID.

If you haven't received any results within a few hours, please contact us and re-submit. See [Known Issues](#). Thank you!

Estimating the Best K

The method of Evanno enables the user to identify a single k value, out of a range of K values, which captures the uppermost level of structure. This method was purposed by Evonno et al. in 2005 (Molecular Ecology 14, 2005). In addition, we also use $\ln(\text{Pr}(X|K))$ values in order to identify the k for which $\text{Pr}(K=k)$ is highest (as described in STRUCTURE's manual, section 5.1).

Run Best K estimation (click for *Instructions*):

Please indicate the format of the uploaded result files:

- ☒ STRUCTURE (.zip) ? [example](#)
- ☐ Log Probability table file ? [example](#)

Upload your file: ? [example](#) Файл не выбран

Email address

Рисунок 20 – Интерфейс онлайн программы CLUMPAK

В файле output.log будут содержаться расчеты показателей для каждого количества кластеров, включая Delta K. Так же можно будет сохранить графический материал.

9. Объединение результатов кластеризации

Применяя алгоритмы популяционно-генетической кластеризации, каждый отдельный генотип разбивается на набор кластеров с использованием стохастических подходов. В результате повторный кластерный анализ одних и тех же данных может дать различные результаты [65]. В процессе кластеризации каждому индивиду присваивается коэффициент или процент «членства» в каждом кластере зависящий от его мультилокусного генотипа. Большая часть различий между повторными анализами объясняется так называемым «переключением меток» (вследствие произвольного маркирования кластеров), однако встречаются случаи, когда результаты значительно отличаются и представляют собой не только перестройку кластеров, но и разницу в их процентном соотношении.

Для решения этих проблем Jakobsson, M. и Rosenberg, N.A. разработали алгоритм оптимально согласующий результаты всех итераций кластерного анализа. Воспользоваться их алгоритмом можно при помощи приложения CLUMPP и онлайн сервиса CLUMPAK

9.1 CLUMPP

CLUMPP (CLUster Matching and Permutation Program) – программа сопоставления и перестановки кластеров. Установочный файл и развернутую инструкцию можно скачать по ссылке (<https://rosenberglab.stanford.edu/clumpp.html>). В программе реализовано три алгоритма оптимального выравнивания результатов всех итераций кластерного анализа, пользователь может самостоятельно выбирать алгоритмы и сравнивать их работу. В качестве входных данных принимается несколько типов файлов, представляющие собой оцененные матрицы с использованием формул Evanno G. (раздел 8). Это файлы формата «INDFILE», в которых указаны

номера образцов, популяций и доля принадлежности к каждому кластеру во всех использованных итерациях (рисунок 21).

1	indnr	(x)	popnr	:	0.315	0.002	0.683
2	indnr	(x)	popnr	:	0.475	0.014	0.511
3	indnr	(x)	popnr	:	0.090	0.005	0.905
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
94	indnr	(x)	popnr	:	0.004	0.003	0.993
95	indnr	(x)	popnr	:	0.004	0.010	0.985
1	indnr	(x)	popnr	:	0.687	0.002	0.310
2	indnr	(x)	popnr	:	0.490	0.011	0.500
3	indnr	(x)	popnr	:	0.898	0.007	0.095
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
94	indnr	(x)	popnr	:	0.993	0.004	0.003
95	indnr	(x)	popnr	:	0.987	0.008	0.005

Рисунок 21 – Пример INDFILE,

где количество образцов равно 95, K=3, итерации – 2

OUTFILE содержит аналогичную матрицу, но с усредненными значениями согласно всем повторениям.

Для того, чтобы воспользоваться программой, необходимо перейти по указанной выше ссылке, выполнить регистрацию и скачать версию, подходящую для вашей платформы. Извлечь полученный архив и скопировать из него 2 файла: приложение CLUMPP.exe и *paramfile*. Скопированные файлы нужно вставить в папку с результатами Structure Harvester (подраздел 8.1), *paramfile* открываем с помощью текстового редактора (например, Блокнот). В открывшемся *paramfile* нужно ввести необходимые нам параметры для анализа:

DATATYPE – в нашем случае значение 0, т.к. для каждого индивида присутствуют отдельные данные;

INDFILE – прописываем имя файла, в котором содержатся индивидуальные данные кластерного анализа, начиная с K2;

POPFIL – имя файла, в котором содержатся данные по группам, в нашем случае удаляем имя;

OUTFILE – файл в котором будут сохранены полученные результаты;

MISCFIL – файл с заданными параметрами (OUTFILE и MISCFIL рекомендуется называть аналогично INDFIL, по соответствующему количеству кластеров);

К – количество кластеров;

С – количество индивидов или популяций;

Р – количество итераций;

М – метод объединения, который вам необходим (как правило FullSearch применяют при небольшом количестве кластеров – до 3-4; Greedy – 5-7; LargeKGreedy – $\geq 6-7$; выбор зависит от объема данных и количества образцов).

Остальные параметры для нашего анализа не требуют изменений, но при необходимости пользователь может их изменить. Сохраняем *paramfile* и запускаем CLUMPP.exe, находящийся в этой же директории. По окончании работы выйдет соответствующее уведомление и будет создано два новых файла с заданными пользователем именами (OUTFILE и MISCFIL), время работы будет зависеть от объема данных, заданных параметров и может занять от нескольких минут до нескольких часов. Далее необходимо настроить *paramfile* для следующего количества кластеров (K3, K4 ...) и последовательно запустить CLUMPP.exe с каждым набором кластеров.

Полученные результаты можно визуализировать с помощью программы «distruct» или «Microsoft Excel». Здесь мы рассмотрим, как это можно сделать в привычном большинству пользователей интерфейсе Microsoft Excel.

Для этого необходимо открыть полученный файл с расширением outfile и скопировать его содержимое на лист Excel, произвести автозамену всех имеющихся знаков «.» на знак «>» и разместить все данные по столбцам. Данные с вероятностями распределения по кластерам используем для создания «Нормированных гистограмм с накоплением» т.к. сумма

вероятностей для каждого индивида будет равняться единице. Далее с помощью настроек можно придать необходимое оформление (рисунок 22).

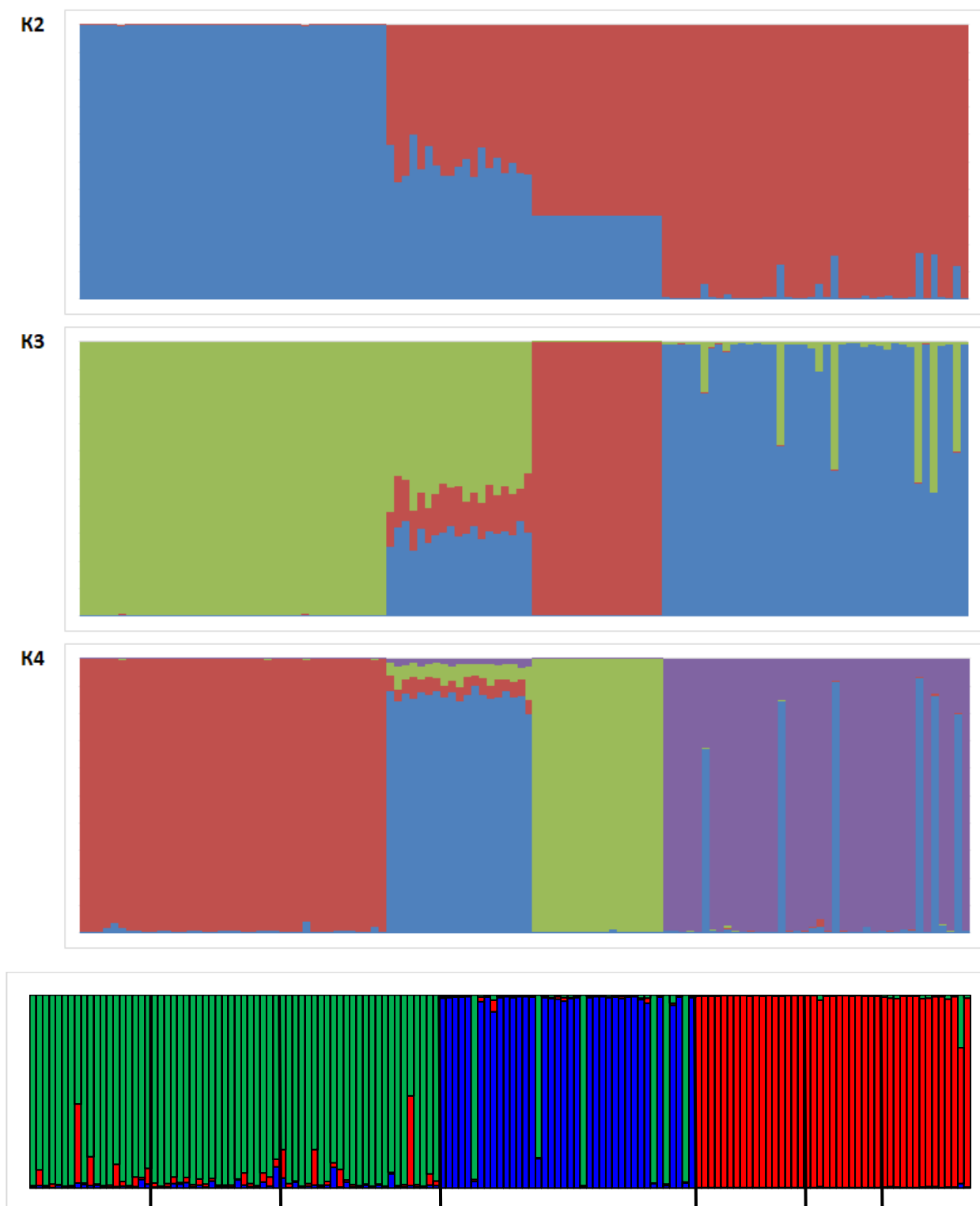
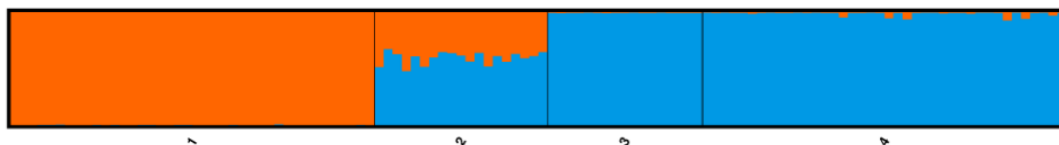


Рисунок 22 – Примеры оформления гистограмм в Microsoft Excel

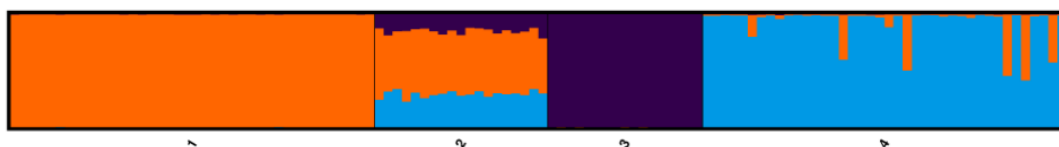
9.2 CLUMPAK

Альтернативой программы CLUMPP выступает онлайн приложение CLUMPAK, ознакомиться с которым можно по ссылке (<https://clumpak.tau.ac.il>) (подраздел 8.2). Для объединения результатов кластеризации нужно использовать вкладку «Main Pipeline», здесь необходимо загрузить архив содержащий результаты работы программы STRUCTURE и запустить работу приложения. По истечении некоторого количества времени приложение выдаст результат, который можно будет сохранить на свой компьютер. Готовый результат будет содержать графический материал в формате pdf, который можно использовать для иллюстрации научных работ, но, к сожалению, редактирование изображения будет невозможно или значительно затруднено (рисунок 23).

K=2



K=3



K=4

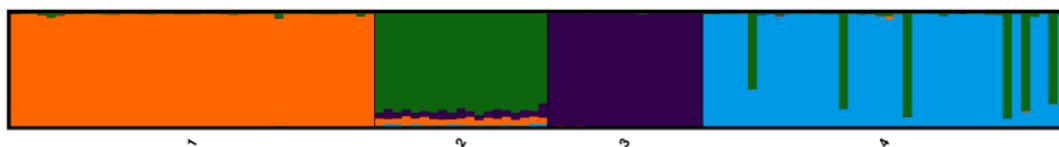


Рисунок 23 – Пример гистограмм, сформированных в приложении CLUMPAK

Таким образом, применяя описанный выше алгоритм, можно подробно описать генетическую структуру популяций, получить для изучаемых выборок большинство из известных показателей генетического разнообразия. Эти данные позволят провести сравнение изучаемых выборок между собой, анализ генетических дистанций, определить генетические процессы, протекающие в исследуемых популяциях. Кластерный анализ дает возможность распределить образцы по генетически однородным группам. С помощью программного обеспечения можно получить достаточно графического материала для написания научных статей и отчетов.

10. Построение филогенетических деревьев в программе SplitsTree4

При анализе генетической структуры, часто возникает необходимость построения филогенетических деревьев для определения генетических дистанций и изучения однородности выборок. Одной из программ, позволяющих построить различные типы дендрограмм, является приложение SplitsTree4 [66, 67].

SplitsTree4 является отличным инструментом, который предоставляет широкие возможности для анализа и визуализации филогенетических данных. Программа обладает интуитивно понятным графическим интерфейсом, поддерживает множество форматов файлов, включая NEXUS, PHYLIP, FASTA и др. Применение программы очень широкое: эволюционная биология

- исследование эволюционных связей между видами, популяционная генетика
- анализ генетического разнообразия и структуры популяций, микробиология
- исследование горизонтального переноса генов и геномной эволюции микроорганизмов, сравнительная геномика - изучение геномных различий и сходств между разными видами.

Как и в других описанных нами программах, для работы SplitsTree4 требуется предоставить данные в поддерживаемом формате. Здесь мы

рассмотрим формат NEXUS (NEXUS file format), используемый в биоинформатике для хранения данных, связанных с филогенетикой, систематикой и популяционной генетикой.

Заголовок всегда состоит из обозначения формата файла – «#NEXUS», далее следуют блоки с параметрами и входными данными (последовательностями, генотипами и т.д.). Каждый блок начинается с ключевого слова «BEGIN» и заканчивается словом «END» (рисунок 24).

```
#NEXUS

BEGIN TAXA;
DIMENSIONS NTAX=227;
END;

BEGIN CHARACTERS;
DIMENSIONS NTAX=227 NCHAR=35
;
  FORMAT
  DATATYPE=STANDARD
  MISSING=?
  LABELS=LEFT
  INTERLEAVE=YES
;
MATRIX
A.baerii_01  00010000000110000000000000000000010
A.baerii_02  0001001000011000000000000000000001000
***
A.baerii_227 0001001000001100000000000000000001000
;
END;
```

Рисунок 24 – Образец данных генотипирования в формате NEXUS

Как видно из рисунка, результаты генотипирования вносятся в виде бинарных матриц, без заголовков, можно также использовать нуклеотидные последовательности. После создания такой матрицы в Microsoft Excel, ее

нужно сохранить в формате .txt с разделением знаками табуляции. В полученном файле прописать два блока: TAXA – количество таксонов; CHARACTERS – где еще раз дублируется количество таксонов, длина последовательности, расположение заголовков образцов и другие параметры.

При открытии сформированного файла через программу формируется дендрограмма построенная на основе введенных данных с использованием алгоритма Neighbor-Net (рисунок 25).

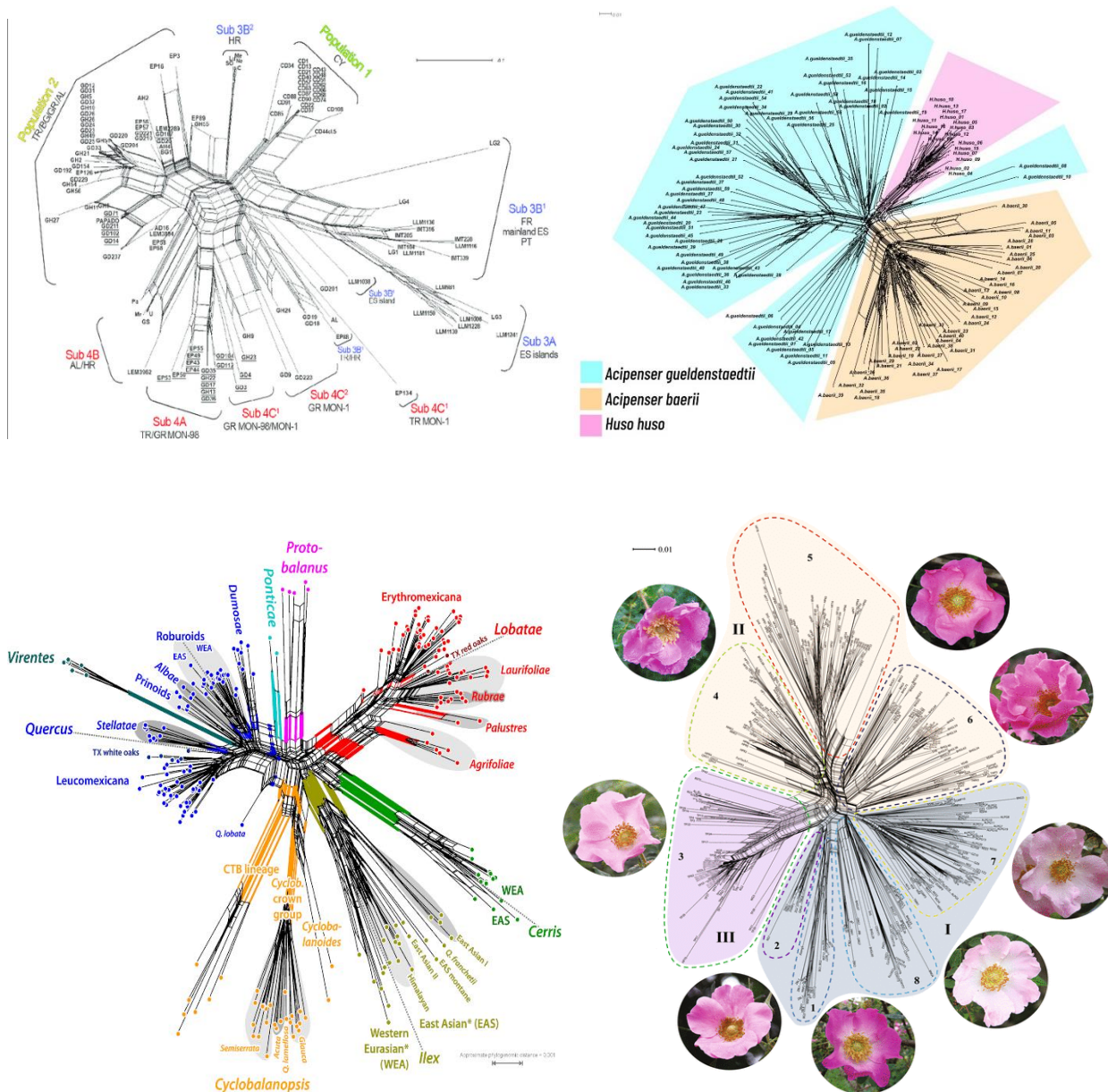


Рисунок 25 – Примеры дендрограмм, построенных с применением алгоритма Neighbor-Net

Пользователю предоставляется широкий выбор настроек и алгоритмов для построения различных видов филогенетических деревьев. Программа позволяет сохранить полученные дендрограммы как в графическом, так и .txt формате.

ЗАКЛЮЧЕНИЕ

Определение генетической структуры популяций осетровых видов рыб и их гибридов представляет собой сложную задачу, требующую комплексного подхода и использования современных методов молекулярной биологии и биоинформатики. В данной рекомендации изложены основные шаги и методы, которые следует учитывать при проведении подобных исследований.

Ключевые выводы:

Выбор маркеров - правильный выбор генетических маркеров играет решающую роль в успешном определении генетической структуры популяций. Современные технологии секвенирования позволяют использовать высокопроизводительные методы для анализа большого числа маркеров.

Методы анализа данных - разнообразие доступных статистических методов и программных пакетов позволяет глубже и точнее анализировать генетические данные, выявляя структуры и паттерны, характерные для различных популяций и гибридов.

Регулярный мониторинг генетического разнообразия популяций осетровых рыб важен для разработки эффективных стратегий их сохранения. Изучение данных о гибридах помогает лучше понять процессы гибридизации и их влияние на популяции.

В подобных исследованиях важно использовать комплексный подход, включающий данные о среде обитания и географическом распространении, который способствует пониманию факторов, влияющих на изменения в генетической структуре популяций.

Результаты анализа состояния генофонда должны быть использованы для разработки программ по сохранению и восстановлению популяций осетровых рыб, управлению рыбными ресурсами, а также для предотвращения нежелательных последствий гибридизации.

В заключение хотелось бы отметить важность применения альтернативных программ при описании генетического разнообразия и сверки

полученных данных. Постоянно развивающаяся сфера биоинформатики приводит к появлению новых концепций и методов расчётов, которые могут значительно отличаться от устоявшихся методик и давать отличные от них результаты. Однако каждый из этих алгоритмов строится на точках зрения авторов приложений и скриптов, что несет в себе определённые риски, но с другой стороны может раскрыть полученных генетические данные с новой стороны.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Международный союз охраны природы (МСОП / IUCN, июль 2021 г.) (<http://www.iucnredlist.org>).
2. Ludwig A. A sturgeon view on conservation genetics // *Eur. J. Wildl. Res.* – 2005. – Vol. 52. – P. 3-8.
3. Peng Z., Ludwig A., Wang D., Diogo R., Wei Q., He S. Age and biogeography of major clades in sturgeons and paddlefishes (Pisces: Acipenseriformes) // *Mol. Phylogenetics Evol.* – 2007. – Vol. 42. – P. 854-862.
4. Fontana F., Congiu L., Mudrak V.A., Quattro J.M., Smith T.I., Ware K., Doroshov S.I. Evidence of hexaploid karyotype in shortnose sturgeon // *Genome.* – 2008. – Vol. 51. – P. 113-119.
5. Butchart S.H., Walpole M., Collen B. et al. Global biodiversity: Indicators of recent declines // *Science.* – 2010. – Vol. 328. – P. 1164-1168.
6. Collen B., Whitton F., Dyer E.E. et al. Global patterns of freshwater species diversity, threat and endemism // *Glob. Ecol. Biogeogr.* – 2014. – Vol. 23. – P. 40-51.
7. Dudgeon D., Arthington A.H., Gessner M.O. et al. Freshwater biodiversity: Importance, threats, status and conservation challenges // *Biol. Rev.* – 2006. – Vol. 81. – P. 163-182.
8. Cincotta R.P., Wisniewski J., Engelman R. Human population in the biodiversity hotspots // *Nature.* – 2000. – Vol. 404. – P. 990-992.
9. Lakra W.S., Sarkar U.K., Kumar R.S. et al. Fish diversity, habitat ecology and their conservation and management issues of a tropical River in Ganga basin, India // *Environmentalist.* – 2010. – Vol. 30. – P. 306-319.
10. Sala O.E., Chapin F.S., Armesto J.J. et al. Global biodiversity scenarios for the year 2100 // *Science.* – 2000. – Vol. 287. – P. 1770-1774.
11. Gehrke P.C., Gilligan D.M., Barwick M. Changes in fish communities of the Shoalhaven River 20 years after construction of Tallowa Dam, Australia // *River Res. Appl.* – 2002. – Vol. 18. – P. 265-286.

12. Meador M.R., Goldstein R.M. Assessing water quality at large geographic scales: Relations among land use, water physicochemistry, riparian condition, and fish community structure // *Environ. Manag.* – 2003. – Vol. 31. – P. 0504-0517.
13. Pof N.L., Allan, J. D. Functional organization of stream fish assemblages in relation to hydrological variability // *Ecology.* – 1995. – Vol. 76. – P. 606-627.
14. Funk W.C., McKay J.K., Hohenlohe P.A., Allendorf F.W. Harnessing genomics for delineating conservation units // *Trends Ecol. Evol.* – 2012. – Vol. 27. – P. 489-496.
15. Harrisson K.A., Pavlova A., Telonis-Scott M., Sunnucks P. Using genomics to characterize evolutionary potential for conservation of wild populations // *Evol. Appl.* – 2014. – Vol. 7. – P. 1008-1025.
16. Moyle L.C., Stinchcombe J.R., Hudgens B.R., Morris W.F. Conservation genetics in the recovery of endangered animal species: A review of US endangered species recovery plans (1977–1998) // *Anim. Biodivers. Conserv.* – 2003. – Vol. 26. – P. 85-95.
17. Pierson J.C., Coates D.J., Oostermeijer J.G.B. et al. Genetic factors in threatened species recovery plans on three continents // *Front. Ecol. Environ.* – 2016. – Vol. 14. – P. 433-440.
18. Frankham R. Genetics and extinction // *Biol. Conserv.* – 2005. – Vol. 126. – P. 131-140.
19. Namkoong G. Genetic conservation strategies: Who needs data? In *Forest Genetics and Sustainability* (ed. Mátyás, Csaba) // Springer. – 1999. – P. 211-216.
20. Pavlova A., Beheregaray L.B., Coleman R. et al. Severe consequences of habitat fragmentation on genetic diversity of an endangered Australian freshwater fish: A call for assisted gene flow // *Evol. Appl.* – 2017. – Vol. 10. – P. 531-550.
21. Taylor H.R., Dussex N., Van Heezik Y. Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners // *Glob. Ecol. Conserv.* – 2017. Vol. 10. – P. 231-242.
22. Frankham R. Effective population size/adult population size ratios in wildlife: A review // *Genet. Res.* – 1995. – Vol. 66. – P. 95-107.

23. Sokal R.R., Jacquez G.M., Wooten M.C. Spatial autocorrelation analysis of migration and selection // *Genetics*. – 1989. – Vol. 121. – P. 845-855.
24. Fluker B.L., Kuhajda B.R., Harris P.M. The effects of riverine impoundment on genetic structure and gene flow in two stream fishes in the Mobile River basin // *Freshw. Biol.* – 2014. Vol. 59. – P. 526-543.
25. Tanaka Y. Extinction of populations by inbreeding depression under stochastic environments // *Popul. Ecol.* – 2000. – Vol. 42. – P. 55-62.
26. Willi Y., Van Buskirk J., Hofmann A.A. Limits to the adaptive potential of small populations // *Annu. Rev. Ecol. Evol. Syst.* – 2006. – Vol. 37. – P. 433-458.
27. Wright L.I., Tregenza T., Hosken D.J. Inbreeding, inbreeding depression and extinction // *Conserv. Genet.* – 2008. – Vol. 9. – P. 833-843.
28. Emadi A., Crim M.T., Brotman D.J. et al. Analytic validity of genetic tests to identify factor V Leiden and prothrombin G20210A // *Am. J. Hematol.* – 2010. – Vol. 85(4). – P. 264-270.
29. Yang W., Kang X., Yang Q., Lin Y., Fang M. Review on the development of genotyping methods for assessing farm animal diversity // *J. Anim. Sci. Bio.* – 2013. – Vol. 4. – P. 1-6.
30. Kumar N.S., Gurusubramanian G. Random amplified polymorphic DNA (RAPD) markers and its applications // *Sci. Vis.* – 2011. – Vol. 11(3). – P. 116-124.
31. Nandani Kumari N., Thakur S.K. Randomly amplified polymorphic DNA- a brief review // *Am. J. of Anim. and Vet. Sci.* – 2014. – Vol. 9(1). – P. 6-13.
32. Weising K., Nybom H., Wolff K., Kahl G. DNA fingerprinting in plants – principles, methods, and applications / 2nd edn. CRC Press, Boca Raton, FL. – 2005. – 472 p.
33. Guimaraes E.P., Ruane J., Scherf B.D., Sonnino A., Dargie J.D. Marker-assisted selection - current status and future perspectives in crops, livestock, forestry and fish / Rome (Food and Agriculture Organization of the United Nations (FAO)) – 2007. – P. 29-49.
34. Brumlop S., Finckh M.R. Applications and potentials of marker assisted selection (MAS) in plant breeding / Final report of the F+E project “Applications

and Potentials of Smart Breeding” (FKZ 350 889 0020) On behalf of the Federal Agency for Nature Conservation. – 2010.

35. Bezerra I.M., Gramacho K.P., Barreto M.A. et al. Genetic diversity and gene flow of the threatened Brazilian endemic parrotfish *Scarus trispinosus* (Valenciennes, 1840) // *Mar. Environ. Res.* – 2018. – Vol. 142. – P. 155-162.

36. Chistiakov D.A., Hellemans B., Volckaert F.A. Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics // *Aquaculture.* – 2006. – Vol. 255. – P. 1-29.

37. Chen L., Bruford M.W., Xu S. et al. Microsatellite variation and significant population genetic structure of endangered finless porpoises (*Neophocaena phocaenoides*) in Chinese coastal waters and the Yangtze River // *Mar. Biol.* – 2010. – Vol. 157. – P. 1453-1462.

38. Goossens B., Chikhi L., Jalil M.F. et al. Patterns of genetic diversity and migration in increasingly fragmented and declining orang-utan (*Pongo pygmaeus*) populations from Sabah, Malaysia // *Mol. Ecol.* – 2005. – Vol. 14. – P. 441-456.

39. Huang X., Börner A., Röder M., Ganai M. Assessing genetic diversity of wheat (*Triticum aestivum* L.) germplasm using microsatellite markers // *Teor. Appl. Genet.* – 2002. – Vol. 105. – P. 699-707.

40. Huey J.A., Balcombe S.R., Real K.M. et al. Genetic structure and effective population size of the most northern population of the Australian river blackfish, *Gadopsis marmoratus* (Richardson 1848): Implications for long-term population viability // *Freshw. Sci.* – 2017. – Vol. 36. – P. 113-123.

41. Liu S., Cantrell R.G., McCarty J.C.Jr., Stewart J.M. Simple sequence repeat-based assessment of genetic diversity in cotton race stock accessions // *Crop Sci.* – 2000. – Vol. 40. – P. 1459-1469.

42. Schwartz M.K., Luikart G., Waples R.S. Genetic monitoring as a promising tool for conservation and management // *Trends Ecol. Evol.* – 2007. – Vol. 22. – P. 25-33.

43. Williamson-Natesan E.G. Comparison of methods for detecting bottlenecks from microsatellite loci // *Conserv. Genet.* – 2005. – Vol. 6. – P. 551-562.

44. Dib C., Fauré S., Fizames et al. C. A comprehensive genetic map of the human genome based on 5,264 microsatellites // *Nature*. – 1996. – V. 380. – P. 152-154.
45. Dietrich W., Miller J., Steen R. A comprehensive genetic map of the mouse genome // *Nature*. – 1996. – Vol.380. – P. 149-152.
46. Gill P., An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes // *Int. J. Legal. Med.* – 2001. – Vol. 114. – P. 204-210.
47. The Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds // *Science*. – 2009. – Vol. 324. – P. 528-532.
48. Brooks S.A., Gabreski N., Miller D., Brisbin A., Brown H.E., Streeter C., Mezey J., Cook D., Antczak D.F. Whole-genome SNP association in the horse: identification of a deletion in myosin Va responsible for lavender foal syndrome. *PLoS Genet.* – 2010. – Vol. 6. – e1000909.
49. Amaral A.J., Megens H.J., Crooijmans R.P.M.A., Heuven H.C.M., Groenen M.A.M. Linkage disequilibrium decay and haplotype block structure in the pig // *Genetics*. – 2008. – Vol. 179. – P. 569-579.
50. Werner M., Sych M., Herbon N., Illig T., König I.R., Wjst M. Large-scaledetermination of SNP allele frequencies in DNA pools using MALDI-TOF mass spectrometry // *Hum. Mutat.* – 2008. – Vol. 20. – P. 57-64.
51. Werner M., Herbon N., Gohlke H., Altmüller J., Knappw M., Heinrich J., Wjst M. Asthma is associated with single-nucleotide polymorphisms in ADAM33 // *Clin. Exp. Allergy*. – 2004. Vol. 34. – P. 26-31.
52. Höss M., Kohn M., Paabo S. et al., Excrement analysis by PCR // *Nature*. – 1992. – Vol. 359. – P. 199.
53. Taberlet P., Bouvet J. Bear conservation genetics // *Nature*. – 1992. – Vol. 358. – P. 197.
54. Morin P.A., Wallis J., Moore J.J. et al. Non-invasive sampling and DNA amplification for paternity exclusion, community structure, and phylogeography in wild chimpanzees // *Primates*. – 1993. – Vol. 34. – P. 347-356.

55. Taberlet P., Waits L.P., Luikart G. Noninvasive genetic sampling: look before you leap // *Trends Ecol. Evol.* – 1999. Vol. 14. – P. 323-327.
56. Henderson-Arzapalo A., King T.L. Novel microsatellite markers for Atlantic sturgeon (*Acipenser oxyrinchus*) population delineation and broodstock management // *Mol. Ecol. Notes.* – 2002. – Vol. 2. – P. 437-439.
57. Welsh A.B., Blumberg M., May B. Identification of microsatellite loci in lake sturgeon, *Acipenser fulvescens*, and their variability in green sturgeon, *A. medirostris* // *Mol. Ecol. Notes.* – 2003. – Vol. 3. – P. 47-55.
58. McQuown E.C., Sloss B.L., Sheehan R.J., et al. Microsatellite analysis of genetic variation in sturgeon: new primers sequences for *Scaphirhynchus* and *Acipenser* // *Trans. Am. Fish. Soc.* – 2000. – Vol. 129. – P. 1380-1388.
59. May B., Krueger C.C., Kincaid H.L. Genetic variation at microsatellite loci in sturgeon: primer sequence homology in *Acipenser* and *Scaphirhynchus* // *Can. J. Fish. Aquat. Sci.* – 1997. – Vol. 54. – P. 1542–1547.
60. Butler J.M. Fundamentals of forensic DNA typing / Academic Press/Elsevier. – 2010. – 519 p.
61. Kim K.S., Sappington T.W. Microsatellite data analysis for population genetics // *Methods Mol. Biol.* – 2013. – Vol. 1006. – P. 271-295.
62. Кластерный анализ // Википедия. [2023]. Дата обновления: 29.06.2023. URL: <https://ru.wikipedia.org/?curid=495696&oldid=131356433> (дата обращения: 03.05.2023).
63. Pritchard J.K., Stephens M., Donnelly P. Inference of population structure using multilocus genotype data // *Genetics.* – 2000. – Vol. 155. – P. 945-959.
64. Evanno G., Regnaut S., Goudet J. Detecting the number of clusters of individuals using the software Structure: A simulation study // *Mol. Ecol.* – 2005. – Vol. 14(8). – P. 2611-2620.
65. Jakobsson M., Rosenberg N.A. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure // *Bioinformatics.* – 2007. – Vol. 23(14). – P. 1801-1806.

66. Bryant D., Moulton V. Neighbor-Net: An agglomerative method for the construction of planar phylogenetic networks // Mol. Biol. Evol. – 2004. – Vol. 21(2). – P. 255-265.

67. Huson D.H., Bryant D. Application of phylogenetic networks in evolutionary studies // Mol. Biol. Evol. – 2006 – Vol. 23(2). – P. 254-267.

**Бейшова И.С., Гинятов Н.С., Ульянов В.А., Ульянова Т.В.,
Ковальчук А.М., Сариев Б.Т., Сидарова А.Ж.**

**Рекомендации
по определению генетической структуры
популяций осетровых видов рыб и их гибридов**



Подписано в печать 25.09.2024 г.
Формат 60х84 1/16 Бумага листовая 80 м/г
Объем 3,19 п.л. Заказ № 235
Тираж 500 экз.

***Отпечатано в полном соответствии
с качеством предоставленных оригиналов***

НАО «Западно-Казахстанский аграрно-технический университет
имени Жангир хана»
090009, г. Уральск, ул. Жангир хана, 51