



Информационно-энтропийный подход к анализу генетического разнообразия популяций (аналитический обзор)

© 2022. В. М. Кузнецов ✉

ФГБНУ «Федеральный аграрный научный центр Северо-Востока имени Н. В. Рудницкого», г. Киров, Российская Федерация

Показано применение информационно-энтропийного анализа на реальных данных: генотипированных по 11 микросателлитным локусам 84 быках джерсейской ($n = 10$), айрширской ($n = 10$), красной датской ($n = 10$), красной шведской ($n = 9$) пород и голштинских экотипов немецкой ($n = 13$), нидерландской ($n = 17$) и североамериканской ($n = 15$) селекций. Информационный индекс Шеннона по объединённым породным выборкам составил 1,695, средневзвешенный по выборкам – 1,325, межвыборочная компонента (S_{HAP}) – 0,370 нит, или 21,8 % (все расчёты по GenAlEx 6.502). Общее абсолютное разнообразие Шеннона составило 5,45, усреднённое по выборкам – 3,76 эффективных аллелей на локус, межпородное – 1,45 эффективных выборов. Относительная оценка дифференциации породных выборов (D'_p) составила 36,2 % и была близка к «традиционным» оценкам по Хедрику, Джосту и Чао. На S_{HAP} -оценки влияния внутривыборочной гетерозиготности не установлено, но отмечена положительная тенденция D'_p -оценок. Матрица парных по породам генетических дистанций (D'_p) и её двумерная проекция имели высокое соответствие с таковыми по $G_{ST(NEI)}$, $F_{ST(W\&C)}$, $G'_{ST(NEI)}$, $F'_{ST(W\&C)}$, D_{JOST} , D_{CHAO} -мерам ($R_{Mantel} \geq 0,94$). Индексы фиксации ($G_{ST(NEI)}$ и $F_{ST(W\&C)}$) следует использовать для изучения эволюционных историй подразделённых популяций. Для анализа современной структуры генетического разнообразия (суб)популяций с акцентом на доминирующие аллели – меры на базе гетерозиготности ($G'_{ST(NEI)}$, $F'_{ST(W\&C)}$, D_{JOST} и D_{CHAO}), при необходимости учитывать все аллели – меру Шеннона (D'_p). В заключение рассматривается семейство степенных мер Хилла порядка $q = 0, 1, 2$ для построения профилей разнообразия, содержащих всю информацию о распределении аллелей в популяции.

Ключевые слова: молочный скот, микросателлиты, дифференциация, генетическая дистанция, информация, энтропия, экспонента Шеннона, анализ главных координат, ординация

Благодарности: работа выполнена при поддержке Минобрнауки РФ в рамках Государственного задания ФГБНУ «Федеральный аграрный научный центр Северо-Востока имени Н. В. Рудницкого» (№ гос. регистрации AAAA-A19-119042290136-4).

Автор благодарит рецензентов за их вклад в экспертную оценку этой работы.

Конфликт интересов: автор заявил об отсутствии конфликта интересов.

Для цитирования: Кузнецов В. М. Информационно-энтропийный подход к анализу генетического разнообразия популяций (аналитический обзор). Аграрная наука Евро-Северо-Востока. 2022;23(2):159-173.

DOI: <https://doi.org/10.30766/2072-9081.2022.23.2.159-173>

Поступила: 27.01.2022

Принята к публикации: 31.03.2022

Опубликована онлайн: 20.04.2022

Information-entropy approach to the analysis of genetic diversity of populations (analytical review)

© 2022. Vasiliiy M. Kuznetsov ✉

Federal Agricultural Research Center of the North-East named N. V. Rudnitsky, Kirov, Russian Federation

The application of information-entropy analysis on real data is shown: 84 bulls of Jersey ($n = 10$), Ayrshire ($n = 10$), red Danish ($n = 10$), red Swedish ($n = 9$) breeds and Holstein ecotypes of German ($n = 13$), Dutch ($n = 17$) and North American ($n = 15$) breeding genotyped by 11 microsatellite loci. The Shannon information index for the combined breed samples was 1.695, the weighted average for the samples was 1.325, the inter-sample component (S_{HAP}) was 0.370 nits or 21.8 % (all calculations according to GenAlEx 6.502). The total absolute diversity of Shannon was 5.45, averaged by samples – 3.76 effective alleles per locus, between breeds – 1.45 effective samples. The relative estimate of the differentiation of breed samples (D'_p) was 36.2 % and was close to the «traditional» estimates for Hedrick, Jost and Cio. The influence of within samples heterozygosity on S_{HAP} estimates has not been established, but a positive trend of D'_p -estimates has been noted. The matrix of paired genetic distances by breeds (D'_p) and its 2D projection had a high correspondence with those by $G_{ST(NEI)}$, $F_{ST(W\&C)}$, $G'_{ST(NEI)}$, $F'_{ST(W\&C)}$, D_{JOST} , D_{CHAO} -measures ($R_{Mantel} \geq 0,94$). Fixation indices ($G_{ST(NEI)}$ and $F_{ST(W\&C)}$) should be used to study the evolutionary stories of subdivided populations. To analyze the current structure of the genetic diversity of (sub)populations with an emphasis on dominant allele measures based on heterozygosity ($G'_{ST(NEI)}$, $F'_{ST(W\&C)}$, D_{JOST} and D_{CHAO}), if necessary, take into account all alleles – the Shannon measure (D'_p). In conclusion, a family of Hill power measures of order $q = 0, 1, 2$ is considered for constructing diversity profiles containing all information about the distribution of alleles in a population.

Keywords: dairy cattle, microsatellites, differentiation, genetic distance, information, entropy, Shannon exponent, analysis of principal coordinates, ordination

Acknowledgements: the research was carried out under the support of the Ministry of Science and Higher Education of the Russian Federation within the state assignment of the Federal Agricultural Research Center of the North-East named N. V. Rudnitsky (theme No. AAAA-A19-119042290136-4).

The author thanks the reviewers for their contribution to the peer review of this work.

Conflict of interest: the author declared no conflict of interest.

For citation: Kuznetsov V. M. Information-entropy approach to the analysis of genetic diversity of populations (analytical review). *Agrarnaya nauka Evro-Severo-Vostoka* = Agricultural Science Euro-North-East. 2022;23(2):159-173. (In Russ.). DOI: <https://doi.org/10.30766/2072-9081.2022.23.2.159-173>

Received: 27.01.2022

Accepted for publication: 31.03.2022

Published online: 24.04.2022

В генетике «разнообразие» – это «набор различий между видами, породами внутри видов и особями внутри пород, выраженными как следствие различий в их ДНК» [1]. Наличие генетического разнообразия и его оценка имеют важное значение для селекции животных, адаптации их к изменяющимся условиям окружающей среды, сохранения исчезающих пород, криоконсервации гамет и эмбрионов.

Разнообразие популяции характеризуют три «экологических» понятия: *богатство* (richness), *обилие* (abundance) и *выравненность* (evenness). Богатство – это число групп (видов, аллелей, морф) в популяции. Обилие – это численность (или относительная частота, доля) каждой морфы. Выравненность – это степень равномерности распределения единиц (наблюдений, животных) популяции по морфам. Меры разнообразия, которые базируются на анализе «богатства», малоинформативны, так как не учитывают «обилие». Популяция, в которой доминируют один-два аллели, генетически менее разнообразна, чем популяция, в которой такое же число равнораспространённых аллелей. Лучшими мерами разнообразия считаются те, которые учитывают богатство, обилие и выравниваемость.

Такие процессы, как изоляция, дрейф генов, отбор и система спаривания приводят к дивергенции популяции, образованию субпопуляций и их дифференциации, в результате чего формируется пространственная структура разнообразия. Разнообразие и дифференциация связаны, т. к. общее разнообразие популяции включает разнообразия внутри субпопуляций и разнообразие между субпопуляциями. Поэтому различают три типа разнообразия: α -разнообразие («в узком смысле») – усреднённое разнообразие отдельных субпопуляций; β -разнообразие (дифференциация) – разнообразие между субпопуляциями, и γ -разнообразие («в широком смысле») – разнообразие в популяции как целое (объединённых субпопуляций).

Показатели разнообразия и дифференциации – важные описательные характеристики популяционной генетики. На уровне маркеров

ДНК генетическое разнообразие измеряется числом аллелей, частотами аллелей, гетерозиготностью. Наиболее широко используется ожидаемая гетерозиготность: $H = 1 - \sum p_m^2$, где p_m – частота m -ой аллели в локусе. Ней назвал H «генным разнообразием», определяемым как вероятность того, что две аллели, выбранных случайным образом из популяции, различны [2]. Если внутрисубпопуляционная гетерозиготность (H_a) гораздо больше, чем межсубпопуляционная (H_b), то говорят о низком уровне дифференциации. Отношение усреднённой внутрисубпопуляционной гетерозиготности (H_a) к гетерозиготности в общей популяции (H_γ) характеризует сходство/родство субпопуляций.

Вместе с тем было показано, что в высокополиморфных системах (например, микросателлиты) гетерозиготность не лучшая мера разнообразия, особенно β -разнообразия [3, 4]. В частности, чем выше H_a , тем ближе H_b к нулю (т. к. $H_b = H_\gamma - H_a$), даже если субпопуляции не имеют общих аллелей. Возрастает вероятность некорректных оценок H_b , ошибочных выводов и неправильных решений, например, при выборе пород(-ы) для сохранения.

Для исследования генетической структуры популяции по ДНК-маркерам имеется два семейства мер [4]. Первое – меры *демографической структуры* или *демографической дифференциации*. Это статистики Райта (F_{ST}), Нея (G_{ST}) [2], Вейра и Кокерхэма ($F_{ST(W\&C)} (= \theta)$) [5]. Их называют «индексами фиксации», т. к. на эмпирическом уровне они, в основном, отражают вероятность фиксации аллелей в каждой субпопуляции, а не фактическую степень дифференциации частот аллелей среди субпопуляций. Второе семейство – меры *аллельной дифференциации* или *структурной дифференциации*. К этому семейству относят модифицированные индексы фиксации ($G''_{ST(HEd)}$, $F'_{ST(W\&C)}$) [6] и меру «истинного» разнообразия Джоста (D_{JOST}) [7] на базе эффективного числа аллелей. Эти меры равны единице, когда каждая субпопуляция состоит исключительно из приватных аллелей, и равны нулю, когда все субпопуляции имеют одни и те же аллели с равными

частотами. Оба семейства мер были рассмотрены нами в работах [8, 9], их применение на реальных данных представлено в [10, 11].

Кроме упомянутых выше метрик, имеются меры разнообразия/дифференциации, «привнесённые» в биологию из статистической физики, теории информации и кибернетики, так называемые «информационно-энтропийные» меры [12, 13]. В теории информации «энтропия» – это мера неопределённости какого-либо опыта, который может иметь разные исходы. Клод Шеннон, один из создателей математической теории информации, в качестве «мер количества информации, возможности выбора и неопределённости» определил величину вида [14]

$$H = -K \sum_{m=1}^M p_m \log p_m = {}^S H,$$

где K – константа (в действительности нужна только для выбора единиц измерения); M – число возможных событий (например, видов в сообществе, аллелей в локусе); p_m – вероятность m -го события из M возможных (в дальнейшем вместо « H » используется « ${}^S H$ », чтобы избежать путаницы с гетерозиготностью¹).

Шеннон писал: «Назовём величину ${}^S H = -\sum p_m \log p_m$ энтропией набора вероятностей $p_1, \dots, p_m$ ». Эта величина есть *усреднённая энтропия* («весами» являются p_m). В то время как величина ${}^S H_m = -\log p_m$ – есть *частная энтропия*, характеризующая только m -ое событие (состояние, исход). Следовательно, энтропия какого-либо опыта (испытания), ${}^S H$, есть усреднённое значение неопределённости отдельных исходов, т. е. случайной величины, принимающей значения $-\log p_m$, с вероятностями p_m . Чем больше вероятность события, тем меньше количество информации в сообщении о таком событии. Получение информации (её увеличение) одновременно означает увеличение знания, что, в свою очередь, означает уменьшение незнания или информационной неопределённости, т. е. энтропии.

Родоначальник кибернетики У. Эшби интерпретировал усреднённую энтропию Шеннона, ${}^S H$, как меру *количества разнообразия*, учитывающую не только абсолютное число разных состояний, но и вероятность, с которой система принимает то или иное состояние [15]. Впоследствии ${}^S H$ даже стали называть «информационный индекс разнообра-

зия Шеннона» или просто «индекс Шеннона», подразумевая разнообразие.

С 1960-х годов ${}^S H$ получил широкое распространение. В последние годы интерес к формуле Шеннона повысился из-за разработки процедуры декомпозиции (разделения) общего индекса (${}^S H_{GT}$) на внутри- и межсубпопуляционные компоненты – ${}^S H_{WP}$ и ${}^S H_{AP}$ соответственно [16]. Кроме того, ${}^S H$, как показатель «количества информации/энтропии», предложили трансформировать через показательную функцию в меру разнообразия – $D = \exp({}^S H)$, где D – является «числовым эквивалентом» («number equivalent») меры разнообразия [17]. Числовой эквивалент – это число одинаково вероятных элементов – эффективных элементов (effective numbers of elements) – необходимых для получения заданного (рассчитанного) значения ${}^S H$ [3]. Так, D_γ (или D_α) – это число *эффективных аллелей* по ${}^S H_{GT}$ (или по ${}^S H_{WP}$), а D_β – число *эффективных субпопуляций* по ${}^S H_{AP}$ (${}^S H_{GT}$, ${}^S H_{WP}$ и ${}^S H_{AP}$ – это индексы Шеннона по объединённой популяции, усреднённый по субпопуляциям и межсубпопуляционный). Если ${}^S H_{GT} = {}^S H_{WP} + {}^S H_{AP}$, то $D_\gamma = D_\alpha \times D_\beta$.

В зарубежных научных публикациях по биологии и экологии D_γ , D_α и D_β используют достаточно широко [18]. В российских исследованиях по генетике и селекции животных, сохранению генофонда исчезающих пород эпизодически используют ${}^S H$ [19, 20]. Вместе с тем считается, что шенноновские меры разнообразия (D_γ , D_α и D_β) имеют некоторые идеальные статистические свойства для измерения биологической информации в различных масштабах [21]. Поэтому *цель статьи* состояла в рассмотрении информационно-энтропийных подходов к анализу разнообразия/дифференциации популяций животных по маркерам ДНК. Ставились задачи: (1) применить информационно-энтропийный анализ к реальным данным – выборкам быков разных пород, генотипированных по микросателлитам; (2) сравнить шенноновские сводные оценки дифференциации пород с таковыми, полученными «традиционными» методами в предыдущих исследованиях [10, 11]; (3) рассчитать матрицу попарных генетических дистанций и на её основе выявить генетическую структуру породных выборок; (4) оценить степень сходства/различия разных метрик разнообразия и визуализировать их ординацию в пространстве малой размерности.

¹Численное значение ${}^S H$ зависит от выбора основания логарифма. При основании 2 единицей ${}^S H$ является *бит*, при основании $e = 2,71828$ – *нит* (*нат*), при основании 10 – *дита* (*хартли*). Так как по определению $p \leq 1$, а логарифм числа меньшего единицы – величина отрицательная, то перед $\log$ находится знак минус.

Материал и методы. Использованы те же данные, что и в предыдущих исследованиях [10, 11]. В частности, 84 быка (=N), каждый генотипирован по 11 STR-локусам (микросателлиты ДНК)², именно: 10 быков джерсейской породы (JER), 10 быков айрширской породы (AYR), 10 – красной датской (RDAT), 9 – красной шведской (RSH) и 45 быков голштинской породы трёх «экотипов»³: 13 быков из Германии (H-DEU), 17 – из Нидерландов (H-NLD), 15 – из США (H-USA).

Анализы информационно-энтропийный и главных координат (Principal Coordinate Analysis, PCoA) были выполнены программой GenAlEx 6.502 [22, 23, 24]. Для регрессионно-корреляционного анализа использовали

программу STATGRAPHICS® Centurion XVI [25]. Результаты сравнивали с таковыми, полученными другими методами в работах [10, 11].

Основная часть. STR-разнообразие породных выборок. В таблице 1 представлены базовые показатели аллельного разнообразия анализируемых породных выборок, которые используются в популяционной генетике. Самый простой – среднее число аллелей на локус – n_a (аллельное богатство). Минимальная величина n_a была в JER-выборке – 3,5 аллели на локус. Породные выборки AYR, RDAT, RSH и H-USA имели около 5 аллелей на локус. Максимальные значения были в H-DEU и H-NLD выборках – около 6 аллелей на локус.

Таблица 1 – Показатели разнообразия породных выборок по STR-маркерам /
Table 1 – Indicators of the diversity of breed samples by STR markers

Порода / Breed	n	n_a	n_e	H_o	H_e	SH	$D (^{S}n_e)$	F	s.e.
JER	10	3,5	2,4	0,56	0,53	0,95	2,6	-0,09	0,074
AYR	10	5,1	3,3	0,69	0,67	1,33	3,8	-0,04	0,059
RDAT	10	5,4	3,2	0,76	0,66	1,32	3,7	-0,15	0,031
RSH	9	4,8	3,4	0,76	0,65	1,30	3,7	-0,15	0,041
H-DEU	13	5,6	3,9	0,74	0,71	1,44	4,2	-0,04	0,062
H-NLD	17	6,2	4,3	0,97	0,75	1,56	4,8	-0,30	0,038
H-USA	15	4,7	3,2	0,70	0,64	1,23	3,4	-0,07	0,040
avEst. (α)	12	5,039	3,371	0,740	0,658	1,304	3,684	-0,123	0,021
totEst. (γ)	84	9,091	4,629	0,756	0,761	1,695	5,447	0,007	0,021

Примечания: n – размер выборки; n_a – число аллелей на локус; n_e – число эффективных аллелей на локус; H_o – наблюдаемая гетерозиготность; H_e – ожидаемая гетерозиготность; SH – индекс Шеннона (нит); D – экспонента SH или эффективное число аллелей на локус по энтропии (Sn_e); F – индекс фиксации; s.e. – стандартная ошибка оценки F; avEst. (α) – среднее по выборкам; totEst. (γ) – оценка по объединённым выборкам /

Notes: n – sample size; n_a – number of alleles per locus; n_e – number of effective alleles per locus; H_o – observed heterozygosity; H_e – expected heterozygosity; SH – Shannon index (nit); D – exponential of SH or the effective number of alleles per locus by entropy (Sn_e); F – fixation index; s.e. – standard error of estimation F; avEst. (α) – average of the samples; totEst. (γ) – estimate based on combined samples

Мера разнообразия n_a не учитывает распространённость аллелей. Поэтому сравнение аллельного разнообразия выборок по n_a корректно только тогда, когда аллели *равночастотны*. Распространённость аллелей, как правило, различная. В таких случаях используют *эффективное число аллелей*, $n_e = 1/\Sigma p_m^2$ [26] (заметим, Σp_m^2 – это гомозиготность по Харди-Вайнбергу; в определении Нея [2] – «генная идентичность»). Величина n_e представляет число *равночастотных* аллелей,

необходимых для получения той же гетерозиготности, что и в фактической выборке [18].

Величина $n_e \leq n_a$, т. к. зависит от высокочастотных (доминирующих) аллелей; чем больше расхождение, тем больше в выборке низкочастотных аллелей. Отношение n_e/n_a характеризует выравненность распространённости аллелей. Минимальная 59,3 % выравненность была в RDAT-выборке, максимальная в RSH – 70,8 %. В усреднённой выборке выравненность составила 66,9 %, в объединённой – 50,9 %.

²STR – Simple Tandem Repeats – участки ДНК длиной 2-6-9 нуклеотидов, tandemно повторенных 5-40 раз. ВНИИплем. База генетических данных быков-производителей. [Электронный ресурс]. URL: <https://www.vniiplem.com/baza-geneticheskikh-dannyh-bykov> (дата обращения: 11.01.2022).

³Экотип – субпопуляция внутри породы, адаптированная к определённой среде обитания.

Для всех породных выборок оценки ожидаемой гетерозиготности (H_e) были на 3-23 % ниже, чем фактической (H_o). В результате индексы фиксации (F) были со знаком «минус», что указывало на преобладание гетерозигот. Однако только в трёх породных выборках (RDAT, RSH и H-NLD) F -оценки были статистически значимыми при $p_{value} < 0,05$ (удвоенные статистические ошибки, s.e., были меньше F -оценок), т. е. имело место достоверное отклонение генотипических частот от равновесия по Харди-Вайнбергу.

На оценки n_e и H_e в большей степени влияют высокочастотные аллели, т.к. квадраты низких частот (p_m^2) есть величины очень малые. Другими словами, при расчёте n_e и H_e более распространённым (доминирующим) аллелям придаётся больший «вес». Поэтому показатели H_e и n_e занижают аллельное разнообразие (суб)популяции, но являются чувствительными индикаторами доминирования (концентрации) одной или нескольких аллелей в локусе(-ах).

Информационно-энтропийный индекс Шеннона, SH , является косвенным показателем разнообразия без чрезмерного акцента на редкие и доминирующие аллели [27, 28], т. к. взвешивает количество энтропии, содержащейся в группе m -ой аллели ($-\ln p_m$) на её частоту (p_m). Низкие значения SH свидетельствуют о малом разнообразии, высокие – о большом. Наименьший индекс Шеннона ($^SH = 0,95$ нит) был отмечен в JER-выборке, наибольший (1,56 нит) – в H-NLD, в остальных выборках на уровне 1,3 нит.

Нижняя граница $^SH = 0$ при $p_m = 1$. В отличие от верхней границы гетерозиготности, равной 1 для любого числа аллелей, максимальная величина SH равна $\ln(M)$, где M – число аллелей в (суб)популяции. Отчасти поэтому SH сложно интерпретировать относительно «количества разнообразия» в (суб)популяциях. Это не означает, что SH является плохим индексом. Напротив, SH является наиболее глубоким и полезным из всех индексов разнообразия, но его величина характеризует энтропию, а не разнообразие [27]. Экспоненциальная функция конвертирует SH в интуитивно интерпретируемый показатель разнообразия – $D = \exp(^SH)$ [17, 18], который Джост [27] определил как *истинное разнообразие*. Применительно к результатам таблицы 1, D (или SH) – это минимальное число необходимых *равночастотных* аллелей для получе-

ния такой же величины SH , которая была рассчитана по реальным выборочным данным [18]. D есть энтропийный аналог эффективного числа аллелей по гетерозиготности (n_e), который называют также «*эффективным числом аллелей по энтропии*». Преобразование SH в D (SH) приводит к легко интерпретируемой, чувствительной мере разнообразия.

Величины SH были больше n_e в среднем на 0,36 аллеля. На одну аллель SH и n_e отличались в AYR, RDAT и H-NLD-выборках. Ранговая корреляция Кендалла (τ) между SH и n_e составила 0,85 ($p_{value} = 0,007$). Для корреляции Кендалла присуща простая вероятностная интерпретация. Так, $\tau = 0,85$ означает, что у 92,5 % пар *порядок* оценок совпадал ($P = (1 + \tau)/2 = (1 + 0,85)/2 = 0,925$), а у 7,5 % – не совпадал.

По трём показателям (n_a , SH и n_e) были построены *профили разнообразия* (рис. 1), которые визуализировали особенности породных выборок. В частности, профиль аллельного разнообразия JER-выборки занимал не только нижний уровень, но и был в некотором удалении от остальных профилей. Кривые профилей разнообразия по всем выборкам достаточно резко снижались, что указывало на значительное число в их аллелофондах низкочастотных аллелей (чем круче – тем больше). Обращает внимание большая *гетерогенность* профилей разнообразия выборок быков голштинских экотипов, в то время как для профилей выборок быков красных пород (пунктирные кривые) характерна *гомогенность* (однородность).

Интересно отметить, что по n_a выборки RDAT и RSH имели ранги 1 и 3 (среди красных пород), по n_e ранги поменялись местами, а по SH аллельное разнообразие выборок было, практически, схожим. Это говорит о том, что, во-первых, в аллелофонде RDAT-выборки низкочастотных аллелей было больше, чем в RSH-выборке, во-вторых, три критерия могут приводить к разным выводам относительно аллельного разнообразия популяций и, соответственно, решениям. Как подтверждение последнего, разнообразие H-USA-выборки по числу фактических аллелей на локус (n_a) было, практически, сходным с таковыми в RSH-выборке. Однако по числу эффективных аллелей на основе энтропии (SH) выборки заметно отличались.

Декомпозиция. По объединённым выборкам информационный индекс Шеннона ($^SH_{GT}$) составил 1,695 нит (табл. 2). В первой

и второй строках даны результаты декомпозиции $^S H_{GT}$ на меж- ($^S H_{AP}$) и внутривыборочный ($^S H_{WP}$) компоненты. Также представлены результаты их конвертирования в «истинное»

разнообразие (D_β , D_α и D_γ). Кроме того, даны относительные оценки разнообразия/дифференциации (D') и перекрытия (O') аллельных профилей.

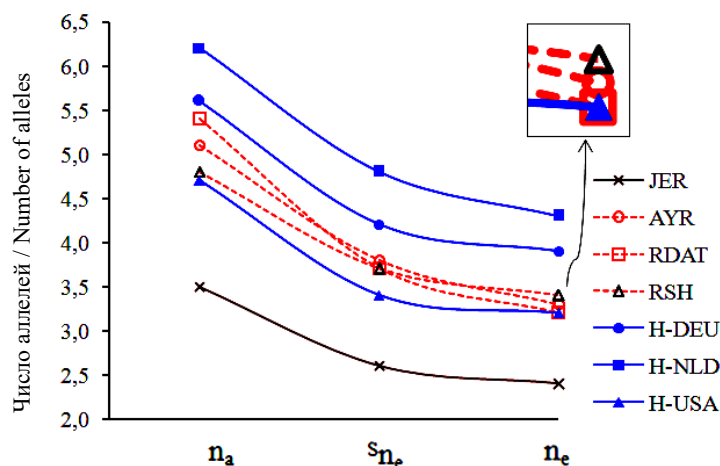


Рис.1 – Профили STR-разнообразия породных выборок. Число аллелей: n_a – фактическое; s_{n_e} – эффективное по энтропии; n_e – эффективное по гетерозиготности (доминирующие) /

Fig.1 – STR-diversity profiles of breed samples. Number of alleles: n_a – actual; s_{n_e} – effective for entropy; n_e – effective for heterozygosity (dominant)

Таблица 2 – Декомпозиция количества информации ($^S H_{GT}$), «истинные» абсолютные (D) и относительные (D') оценки разнообразия и перекрытия (O') аллельных профилей породных выборок /

Table 2 – Decomposition of the amount of information ($^S H_{GT}$); «true» absolute (D) and relative (D') estimates of diversity and overlap (O') of allelic profiles of breed samples

Source Inform.	$^S H$	$^S H, \%$	$D = \exp(^S H)$	$D', \%$	$O', \%$	p_{perm}
Among Pops, AP – (β)	0,370	21,8	1,45	36,2	63,8	0,001
Within Pops, WP – (α)	1,325	78,2	3,76	76,5	23,5	1,000
Total, GT – (γ)	1,695	100,0	5,45	82,1	17,9	-

Примечания: $O' = 100 - D'$. p_{perm} – достигнутый уровень статистической значимости (пермутационный тест с 999 рандомизированными перестановками) /

Notes: $O' = 100 - D'$. p_{perm} – achieved level of statistical significance (permutation test with 999 random shuffles)

Межвыборочная компонента ($^S H_{AP}$) оценивалась в 0,370 нит, внутривыборочная ($^S H_{WP}$) – 1,325 нит. Так как для информационной меры Шеннона характерна аддитивная структура ($^S H_{GT} = ^S H_{AP} + ^S H_{WP}$), то доля межвыборочной компоненты составила $^S H_{AP} / ^S H_{GT} = (0,370/1,695) = 0,218$, или 21,8 %. Это относительный показатель количества межвыборочной информации, но не разнообразия аллельных профилей выборок. Показательная функция перевела $^S H_{GT}$, $^S H_{WP}$ и $^S H_{AP}$ в меры разнообразия (D) с мультипликативным разложением:

- $^S H_{GT} = 1,695$ нит в $D_{GT} = \exp(^S H_{GT}) = 5,45$ – число эффективных аллелей по энтропии в объединённой выборке ($=D_\gamma$);

- $^S H_{WP} = 1,325$ нит в $D_{WP} = \exp(^S H_{WP}) = 3,76$ – число эффективных аллелей по энтропии в усреднённой выборке ($=D_\alpha$);

- $^S H_{AP} = 0,370$ нит в $D_{AP} = \exp(^S H_{AP}) = 1,45$ – эффективное число породных выборок ($=D_\beta$).

Последнее можно получить из отношения D_γ/D_α ($D_\beta = 5,45/3,76 = 1,45$), т. к.

$$D_\gamma = D_\alpha \times D_\beta (= 3,76 \times 1,45 = 5,45).$$

D_β – это абсолютная мера дифференциации субпопуляций. Если имеется P субпопуляций равной численности и нет общих аллелей, то D_β всегда будет равно P (100% дифференциация), независимо от частоты аллелей. Если выборки имеют одни и те же аллели с одинаковыми частотами, то D_β будет равно 1 (нет дифференциации – все субпопуляции являются частями одной популяции). По имеющимся данным величина абсолютной дифференциации составила, примерно, 1,5 эффективной породной выборки (фактически – 7 выборок). В относительных единицах $D'_\beta = 36,2 \%$,

т. е. по частотам аллельные профили выборок различались в среднем на 36,2 % и на 63,8 % ($100 - D'_\beta = O'_\beta$) перекрывались. В усреднённой породной выборке гетерозиготность по энтропии составила 76,5 %, общая по объединённым выборкам – 82,1 % (соответствующие H_e были 0,658 и 0,761 (табл. 1)).

Эффект внутривыборочной гетерозиготности. В предыдущих исследованиях [10, 11] на тех же STR-данных было показано, что высокая внутривыборочная гетерозиготность (cH_s), характерная для микросателлитов, являлась причиной получения существенно заниженных оценок β -разнообразия методами Нея и AMOVA⁴. Была проведена проверка влияния уровня cH_s на шенноновские оценки ($^sH_{AP}$, D'_β и D'_β). Для этого имеющийся набор данных был разделён на 11 псевдонаборов –

каждый с разным локусом. По этим псевдонаборам были рассчитаны cH_s (а также n_a и n_e) и оценки $^sH_{AP}$, D'_β и D'_β . Затем была сформирована выборка с шестью переменными (n_a , n_e , cH_s , $^sH_{AP}$, D'_β и D'_β) и 11 записями, по которой были проведены регрессионно-корреляционные анализы (рис. 2).

Разброс полокусных оценок $^sH_{AP}$ (рис. 2, слева) и линия регрессии свидетельствовали об их [оценках] независимости от уровня внутривыборочной гетерозиготности. Это подтверждал и нулевой коэффициент детерминации ($R^2 = 0,001$). В то же время распределение оценок D'_β (рис. 2, справа) указывало на наличие положительной регрессии. Однако коэффициент детерминации этой тенденции был недостаточно высоким ($R^2 = 0,55$).

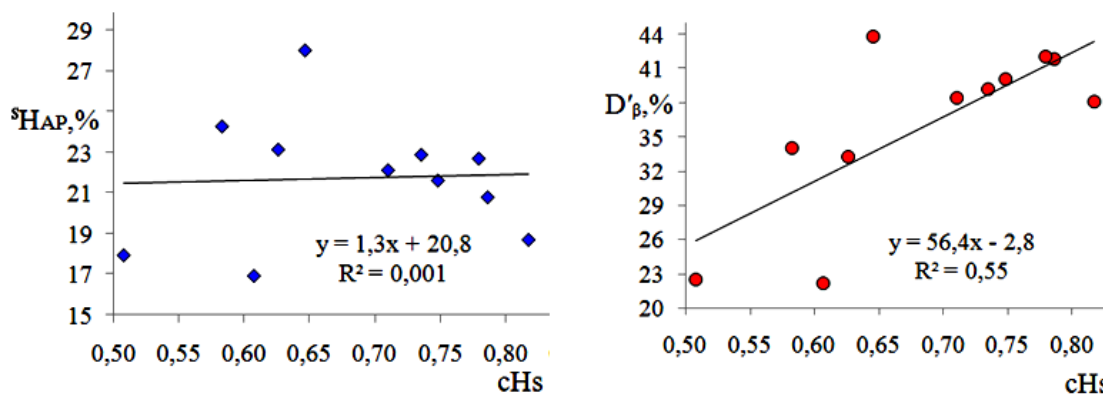


Рис. 2. Распределение полокусных оценок $^sH_{AP}$ и D'_β в зависимости от внутривыборочной гетерозиготности (cH_s). Распределение D'_β было аналогично распределению D'_β /

Fig. 2. Distribution of polocus estimates of $^sH_{AP}$ and D'_β (estimates for each locus) depending on intrasample heterozygosity (cH_s). The distribution of D'_β was similar to the distribution of D'_β

Распределение D'_β было аналогично распределению D'_β . Взаимосвязь D'_β и D'_β оценок с числом аллелей на локус (n_a), как и с числом эффективных аллелей (n_e), также была схожей. Коэффициенты корреляции по Пирхнеру составляли: с n_a – 0,64 ($p_{value} = 0,031$), с n_e – 0,75 ($p_{value} = 0,007$); корреляции по Кендаллу были 0,41 ($p_{value} = 0,078$) и 0,52 ($p_{value} = 0,025$) соответственно.

Сравнение с оценками по другим методам. Было проведено сопоставление показателей β -разнообразия породных выборок, полученных в модуле Shennon программы GenAlEx ($^sH_{AP}$ и D'_β), с таковыми, рассчитанными ранее

в модуле G-Statistics (по Нею – G_{ST} , с поправкой Хедрика на гетерозиготность – $G''_{ST(HED)}$) и по Джосту – D_{JOST}), в модуле AMOVA (по Вейру и Кокерхэму – $F_{ST(W\&C)}$, с поправкой на гетерозиготность – $F'_{ST(W\&C)}$), а также с оценкой методом Чао-Морисита-Хорна, рассчитанной в программе SpadeR (D_{CHAO}) [29]. Результаты представлены в таблице 3.

Оценка разнообразия Шеннона ($D'_\beta = 36,2$ %) была близка к скорректированным на внутривыборочную гетерозиготность оценкам по Нею и AMOVA ($G''_{ST(HED)} = 37,5$ % и $F'_{ST(W\&C)} = 35,1$ %) и на 6,6 процентных пункта превышала оценки аллельной дифференциации по Джосту и Чао

⁴Заметим, при анализе выборок по двуаллельным SNP'ам (однонуклеотидный полиморфизм) меры дифференциации, основанные на гетерозиготности, корректны только в случае сравнения двух выборок, или если несколько выборок анализируются попарно [4].

($D_{JOST} = 29,2\%$ и $D_{CHAO} = 30,8\%$). Информационная мера Шеннона ($S_{HAP} = 21,8\%$) занимала промежуточное положение между оценками аллельной дифференциации (D_{JOST} , D_{CHAO}) и нескорректированными индексами фиксации (G_{ST} , $F_{ST(W\&C)}$). Последние были на

уровне 10 % и характеризовали не уровень аллельной дифференциации породных выборок, а степень приближения к состоянию полной фиксации аллелей, когда каждое животное является гомозиготным по исследуемому локусам [30].

Таблица 3 – Оценки β -разнообразия, рассчитанные разными методами по одним и тем же STR-данным / Table 3 – Estimates of β -diversity, calculated by different methods using the same STR data

Мера / Measure	Est., %	p_{perm}	95CI _U	95CI _L	Источник / Source
$G_{ST(NEI)}$	10,3	0,001	8,0	13,0	[10]
$F_{ST(W\&C)}$	10,8	0,001	8,6	13,3	[11]
$G''_{ST(HED)}$	37,5	0,001	32,4	42,9	[10]
$F'_{ST(W\&C)}$	35,1	0,001	28,0	43,0	[11]
D_{JOST}	29,2	0,001	24,4	33,9	[10]
D_{CHAO}	30,8	-	27,5	34,2	[11]
S_{HAP}	21,8	0,001	-	-	Table 2
D'_{β}	36,2	0,001	-	-	Table 2

Примечания: Est. – оценка; 95 %; CI – 95 % доверительный интервал; субиндексы L и U – нижняя и верхняя границы CI /

Notes: Est. – estimate; 95 %; CI – 95 % confidence interval; subindexes L and U – lower and upper bounds of CI.

Коэффициенты корреляции Пирхнера полокусных D'_{β} -оценок были статистически значимыми только с D_{JOST} - и D_{CHAO} -оценками ($r = 0,6$; $p_{value} = 0,049$). Ранговая корреляция Кендалла составляла 0,35 ($p_{value} = 0,14$) и информировала о том, что с вероятностью 86 % совпадение рангов оценок возможно у $\frac{2}{3}$ пар.

Ординация породных выборок. Относительная оценка β -разнообразия Шеннона ($D'_{\beta} = 36,2\%$) характеризовала обобщённую аллельную дифференциацию породных выборок, нивелируя двухсторонние генетические отношения. В таблице 4 представлена матрица генетических дистанций (Genetic Distance, GD) между аллельными профилями всех попарных комбинаций семи породных выборок.

Таблица 4 – Треугольная матрица парных генетических дистанций (под диагональю – D'_{β} -оценки, над диагональю – p_{perm}) /

Table 4 – Triangular matrix of paired genetic distances (under the diagonal – D'_{β} -estimates, above the diagonal – p_{perm})

Порода / Breed	JER	AYR	RDAT	RSH	H-DEU	H-NLD	H-USA
JER	-	0,001	0,001	0,001	0,001	0,001	0,001
AYR	0,502	-	0,227	0,203	0,001	0,001	0,001
RDAT	0,537	0,180	-	0,110	0,001	0,001	0,001
RSH	0,593	0,169	0,198	-	0,001	0,001	0,001
H-DEU	0,508	0,422	0,420	0,419	-	0,168	0,286
H-NLD	0,460	0,379	0,404	0,408	0,120	-	0,001
H-USA	0,526	0,439	0,402	0,426	0,115	0,179	-

С вероятностью ошибки $\alpha_{Bonf} = 0,0024$ (множественный тест Бонферрони) из 21 парного сравнения 16 оценок GD были статистически значимыми ($p_{perm} < \alpha_{Bonf} = 0,0024$). Оцен-

ки GD между джерсейской (JER) и остальными породными выборками были в пределах 50-60 % ($p_{perm} < 0,001$). GD среди красных пород (AYR, RDAT и RSH) были на уровне 20 %, GD между джерсейской (JER) и остальными породными выборками были в пределах 50-60 % ($p_{perm} < 0,001$). GD среди красных пород (AYR, RDAT и RSH) были на уровне 20 %, GD между джерсейской (JER) и остальными породными выборками были в пределах 50-60 % ($p_{perm} < 0,001$).

но статистически *незначимые*. Среди голштинских экотипов (H-DEU, H-NLD и H-USA) статистически значимая была только GD между H-NLD и H-USA выборками ($D'_\beta = 18\%$). Между красными породами и голштинскими экотипами GD были на уровне 40 % ($p_{perm} < 0,001$)⁵.

Для выявления структурной дифференциации пород в пространстве был использован анализ главных координат (Principal Coordinate Analysis, PCoA). Метод PCoA проецирует данные матрицы GD между породами в достаточно адекватный 2D (3D) график *ординации* – взаимном расположении пород при минимально возможном искажении расстояния между ними. Скрытая в D'_β -матрице пространственная генетическая структура породных выборок визуализирована на рисунке 3.

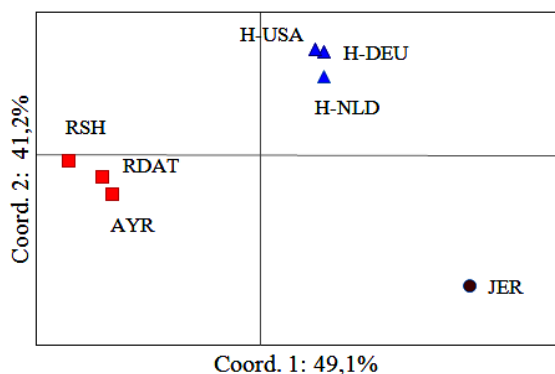


Рис. 3. Ординация породных выборок в двумерном пространстве (PCoA по D'_β -матрице) /

Fig. 3. Ordination of the breed samples in two-dimensional space (PCoA by the D'_β -matrix)

Две координаты суммарно объясняли 90,3 % общей дисперсии, заключённой в матрице GD. Следовательно, двумерное решение было адекватным, а информативность PCoA достаточно высокой (потеря информации 9,7 %). На 2D диаграмме PCoA выделил кластер «красных» пород (RED), кластер голштинских «экотипов» (HOL) и ветку джерсейской породы (JER).

Малые статистически незначимые (5 из 6) GD между породными выборками в пределах кластеров свидетельствовали о большом сходстве аллельных профилей и возможности рассмотрения их, как две «породы»: RED и HOL. Сводная оценка аллельной дифференциации укрупнённых выборок (JER, RED и HOL) составила $D'_\beta = 38,6\%$ с $p_{perm} < 0,001$ (на 7 % больше, чем по семи породным выборкам). Соответствующие сводные оценки по AMOVA ($F'_{ST(W\&C)}$) и D_{JOST} были 45,3 и 42,7 % [11]. Ниже представлены парные по укрупнённым породным выборкам GD (%):

GD	D'_β	$F'_{ST(W\&C)}$ [11]	D_{JOST} [11]
JER - RED	45	65	55
JER - HOL	42	51	41
RED - HOL	34	37	29.

Генетические дистанции, рассчитанные разными методами, были достаточно близки и имели общую тенденцию: $GD_{JER-RED} > GD_{JER-HOL} > GD_{RED-HOL}$ (большая вероятность перекрытия интервальных оценок).

Данные по группам RED и HOL (регионам, Regs) были подвергнуты двухуровневой декомпозиции, результаты которой представлены в таблице 5.

Межрегиональная компонента (Among Regs) информации (S_{HAR} , %) составила 11,0 %, разнообразия Шеннона (D'_β) – 34,3 %. Последняя хорошо соответствовала оценке GD между «породами» RED и HOL – 34 %. Аллельная дифференциация породных выборок в пределах регионов была в 2,2 раза ниже, чем «регионов» ($D'_\beta = 15,9\%$). Перекрытие частот аллелей (общих аллелей) в породных выборках (O'_β) было 84,1 %, в регионах (O'_δ) – 65,7 %. Уровни гетерозиготности по энтропии ($D'_\alpha = 77,8\%$ и $D'_\gamma = 81,7\%$) незначительно отличались от таковых в таблице 2 (76,5 и 82,1 %).

⁵Для сравнения: GD современной голштинской породы с локальными, архивными и древними образцами ДНК российских пород были в разы меньше. Например, с черно-пестрой породой (1970-1980-х гг.) – 8,4 % [31], с пещёрским скотом – 4,1 % [32], с ярославской породой и великорусским скотом (конец XIX начало XX веков) – 6,2 и 1,6 % [33]. Получается, что русский обыкновенный (великорусский) скот (по акад. А. Ф. Миддендорфу – «первичнолесной», «горемычки» и «тосканки» (Миддендорф А. Ф. Отчет о породе крупного рогатого скота Северной России и улучшении его. Исследование современного состояния скотоводства в России: Рогатый скот. М., 1884-1885); по проф. Н. П. Чирвинскому – «северный»; по К. С. Трипольскому – «беспородистый» [34]) на 98,4 % был «genetic relationship» голштинской породе, которая десятилетиями интенсивно селекционировалась, т. е. «горемычки» и голштинки – члены одной популяции!? Отметим, что независимо от меры разнообразия/дифференциации, если выборки небольшие, то некоторые редкие аллели (суб)популяций будут в них отсутствовать. Это означает, что выборочные частоты будут систематически завышать фактические частоты аллелей в (суб)популяциях и, следовательно, недооценивать разнообразие и смещать оценки дифференциации.

Таблица 5 – Двухуровневая декомпозиция информационной меры Шеннона /
Table 5 – Two-level decomposition of the Shannon information measure

Источник / Source Inform.	SH	$^SH, \%$	$D = \exp(^SH)$	$D', \%$	$O', \%$	P_{perm}
Among Regs, AR – (δ)	0,183	11,0	1,20	34,3	65,7	0,001
Among Pops, AP – (β)	0,112	6,6	1,12	15,9	84,1	0,001
Within Pops, WP – (α)	1,376	82,4	3,96	77,8	22,2	1,000
Total, GT – (γ)	1,670	100,0	5,31	81,7	18,3	-

Ординация мер дифференциации. Пространственная структура аллельного разнообразия породных выборок на рисунке 3 визу-

ально была очень схожа с таковыми по матрицам иных мер дифференциации [10, 11], хотя оценки GD существенно отличались (табл. 6).

Таблица 6 – Треугольные матрицы парных генетических дистанций, рассчитанных семью методами (табличный формат) /

Table 6 – Triangular matrixes of paired genetic distances, calculated by seven methods (table format)

BR1	BR2	$G_{ST(NEI)}$	$F_{ST(W\&C)}$	$G''_{ST(HED)}$	$F'_{ST(W\&C)}$	D_{JOST}	D_{CHAO}	D'_{β}
JER	AYR	0,122	0,217	0,590	0,590	0,477	0,447	0,502
JER	RDAT	0,143	0,250	0,663	0,663	0,551	0,510	0,537
AYR	RDAT	0,001	0,001	0,005	0,005	0,003	0,034	0,180
JER	RSH	0,159	0,275	0,728	0,729	0,626	0,558	0,593
AYR	RSH	0,004	0,008	0,026	0,026	0,019	0,040	0,169
RDAT	RSH	0,003	0,005	0,017	0,017	0,012	0,027	0,198
JER	H-DEU	0,102	0,181	0,524	0,522	0,417	0,439	0,508
AYR	H-DEU	0,060	0,112	0,401	0,401	0,325	0,339	0,422
RDAT	H-DEU	0,064	0,120	0,417	0,416	0,337	0,321	0,420
RSH	H-DEU	0,064	0,119	0,416	0,414	0,336	0,315	0,419
JER	H-NLD	0,092	0,163	0,501	0,498	0,401	0,408	0,460
AYR	H-NLD	0,046	0,089	0,335	0,337	0,270	0,286	0,379
RDAT	H-NLD	0,053	0,100	0,371	0,371	0,300	0,300	0,404
RSH	H-NLD	0,056	0,105	0,388	0,388	0,316	0,305	0,408
H-DEU	H-NLD	0,004	0,008	0,032	0,034	0,024	0,037	0,120
JER	H-USA	0,123	0,216	0,565	0,563	0,443	0,472	0,526
AYR	H-USA	0,065	0,124	0,387	0,389	0,302	0,314	0,439
RDAT	H-USA	0,065	0,123	0,377	0,377	0,289	0,292	0,402
RSH	H-USA	0,064	0,121	0,372	0,372	0,285	0,288	0,426
H-DEU	H-USA	0,003	0,006	0,019	0,020	0,014	0,032	0,115
H-NLD	H-USA	0,015	0,030	0,107	0,107	0,079	0,086	0,179

Примечания: BR – порода. На $G_{ST(NEI)}$ и $F_{ST(W\&C)}$ оценки влиял уровень гетерозиготности выборок. В $G''_{ST(HED)}$ и $F'_{ST(W\&C)}$ оценках влияние гетерозиготности устранено. На D-оценки уровень гетерозиготности выборок не оказывает влияние /

Notes: BR – breed. The $G_{ST(NEI)}$ and $F_{ST(W\&C)}$ estimates were influenced by the level of heterozygosity of the samples. In the $G''_{ST(HED)}$ and $F'_{ST(W\&C)}$ estimates, the influence of heterozygosity is eliminated. The level of heterozygosity of the samples does not affect the D-estimates

Информационная мера Шеннона непредвзято учитывала вклад низкочастотных аллелей. Возможно, именно этим свойством объясняется то, что GD по D'_{β} -мере между породными выборками AYR и RDAT, AYR и RSH, RDAT и RSH, H-DEU и H-NLD, H-DEU и H-USA

были в разы больше, чем по остальным мерам аллельной дифференциации.

Для оценки сходства матриц GD в таблице 6 использован тест Мантеля (Mantel correlation, R_M). Результаты представлены в таблице 7.

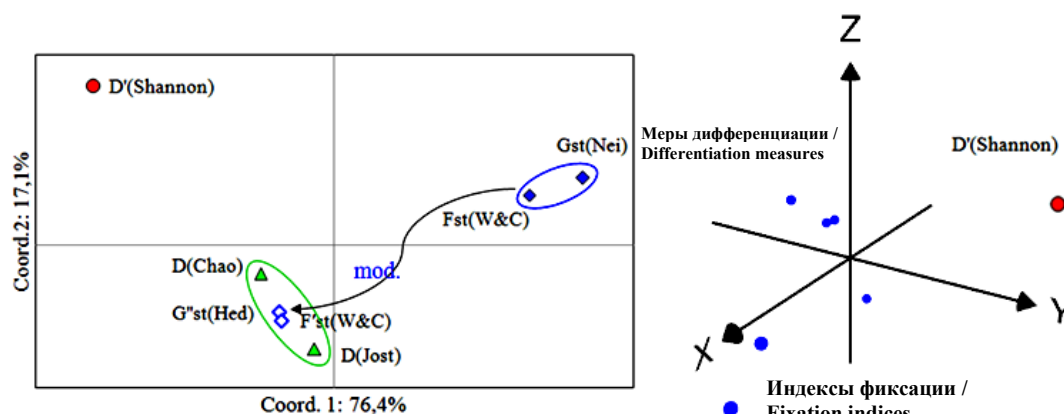
Таблица 7 – Корреляции Мантеля между матрицами генетических дистанций, рассчитанных разными методами (сходство матриц) /

Table 7 – Mantel correlations between genetic distance matrixes calculated by different methods (similarity of matrixes)

Мера / Measure	$G_{ST(NEI)}$	$G''_{ST(HEd)}$	$F_{ST(W\&C)}$	$F'_{ST(W\&C)}$	D_{JOST}	D_{CHAO}
$G''_{ST(HEd)}$	0,973	-	-	-	-	-
$F_{ST(W\&C)}$	0,999	0,982	-	-	-	-
$F'_{ST(W\&C)}$	0,973	1,000	0,982	-	-	-
D_{JOST}	0,976	0,998	0,983	0,998	-	-
D_{CHAO}	0,971	0,997	0,978	0,997	0,994	-
D'_β (Shannon)	0,938	0,982	0,950	0,981	0,975	0,984

Коэффициенты R_M были в диапазоне $0,938 \div 1,0$. Их большое число и малые различия делали проблематичным выявление каких-либо структурных отношений между мерами. Поэтому были привлечены методы многомерной статистики для визуализации возможной структуры. Предварительно коэффициенты R_M матрицы сходства (табл. 7) были трансформированы в коэффициенты «несходства» ($= 1 - R_M$), чтобы сформировать матрицу дистанций между метриками (не приводится). Её проекции представлены на рисунке 4 в двумерной плоскости (слева; модуль PCoA программы GenAlEx) и в трёхмерном пространстве (справа; программа Graphing Calculator 3D).

Информативность PCoA составила $76,4 + 17,1 = 93,5\%$ (потеря информации 6,5 %). Первая координата отделила семейство мер индексов фиксации ($G_{ST(NEI)}$, $F_{ST(W\&C)}$) от семейства мер аллельной дифференциации. Вторая координата выделила методы аллельной дифференциации на основе гетерозиготности (D_{JOST} , D_{CHAO} , $G''_{ST(HEd)}$ и $F'_{ST(W\&C)}$). При этом $G''_{ST(HEd)}$ - и $F'_{ST(W\&C)}$ -меры – модифицированные индексы фиксации (устранено влияние уровня внутривыборочной гетерозиготности). В общем, PCoA-процедура сформировала кластер индексов фиксации $\{G_{ST(NEI)}, F_{ST(W\&C)}\}$, кластер мер аллельной дифференциации $\{D_{JOST}, D_{CHAO}$ и $G''_{ST(HEd)}, F'_{ST(W\&C)}\}$ и отдельно выделила информационную меру β -разнообразия Шеннона $\{D'(\text{Shannon})\}$.

Рис. 4. 2D и 3D ординации семи мер популяционной дифференциации (mod. – модификация $G_{ST(NEI)}$ и $F_{ST(W\&C)}$) /Fig. 4. 2D and 3D ordinations of seven measures of population differentiation (mod. – modification of $G_{ST(NEI)}$ and $F_{ST(W\&C)}$)

Небольшие расстояния между мерами внутри кластеров указывали на их подобие; большие расстояния между кластерами (включая «ветку») свидетельствовали об их отличии. По аналогии с «генетическими», расстояния

между точками-мерами можно было бы назвать «методическими». Но, например, несмотря на большие различия в методиках расчёта $G_{ST(NEI)}$ и $F_{ST(W\&C)}$, эти метрики образовали общий кластер. Поэтому в данном случае

для расстояний/дистанций, возможно, больше подходит определение «концептуальные». И этому имеется некоторое обоснование, именно: в основе мер $G_{ST(NEI)}$ и $F_{ST(W\&C)}$ – концепция фиксации аллелей, меры D_{JOST} , D_{CHAO} , $G'_{ST(NEI)}$ и $F'_{ST(W\&C)}$ базируются, в общем, на гетерозиготности, D'_β -мера – на теории информации. Концептуальные различия метрик выразились в коэффициентах матрицы несходства и через PCoA-процедуру визуализировались в их ординации, которая, на наш взгляд, представляется вполне логичной.

3D-проекция (рис. 4, справа) дополнительно учла 4,4 % дисперсии (ось Z) матрицы несходства и сократила потерю информации до $100 - (93,5 + 4,4) = 2,1$ %. Но, как и следовало ожидать, это существенно не отразилось на ординации исследуемых метрик.

Степенные меры разнообразия. Хиллом [13] в 1973 году была представлена *единая концепция разнообразия* с семейством *степенных мер* вида (Hill's numbers – «числа Хилла»):

$${}_qD \equiv \left[\sum_{m=1}^M p_m^q \right]^{1/(1-q)} = ({}^q\lambda)^{1/(1-q)}.$$

${}_qD$ называют «*истинным разнообразием порядка q*». Показатель степени «q» в правой части тождества ($\equiv$) есть «порядок» меры разнообразия. Степенные меры производят оценки *эффективного числа элементов*, в частности, аллелей и отличаются только своей возможностью учитывать или игнорировать редкие аллели. По мере увеличения порядка q меры ${}_qD$ становятся всё более чувствительными к частотам аллелей. Так, при $q = 0$ мера ${}_0D$ игнорирует распространённость (частоту) аллелей и выражает разнообразие через аллельное богатство. Значения q между нулём и единицей придают в ${}_qD$ больший вес редким аллелям. Когда $q = 1$, то мера ${}_1D$ уравнивает вклады богатства и распространённости. Порядки выше 1 придают больший вес наиболее распространённым аллелям. При $q = 2$ мера ${}_2D$ подчёркивает выравненность аллелей больше, чем богатство, и равна эффективному числу аллелей по гетерозиготности.

Критическая точка, которая взвешивает все аллели по их частотам (не отдавая предпочтения ни доминирующим, ни редким аллелям), при $q = 1$. ${}_qD$ при $q = 1$ не определено, но его предел существует и равен экспоненте энтропии Шеннона – ${}_1D = \exp({}^SH)$! Важно, мера ${}_1D$ была получена Хиллом не из информационной теории, а выведена естественным путём из формализации числовых эквивалентов [27].

Мера ${}_1D$ – это «эффективное число аллелей по энтропии» (в табл. 1 как D или Sn_e). ${}_1D$ наиболее подходящая мера γ -разнообразия. ${}_1Du$ также единственное «истинное» разнообразие, которое можно разложить на независимые α - и β -разнообразия.

Мера ${}_1D$ – противоположность мерам, основанным на гетерозиготности, которые имеют порядок $q = 2$ и придают непропорционально больший вес более распространённым аллелям. Поэтому, если интерес представляют только доминирующие аллели, то для оценки разнообразия следует использовать меру с $q = 2$; в противном случае меру с $q = 1$. В то же время имеется мнение [4, 35], что необходимо использовать несколько мер одновременно для оценивания различных аспектов разнообразия, к которым каждая мера наиболее чувствительна (рис. 1 в качестве примера). Считается полезным использовать числа Хилла порядка $q = 0, 1, 2$, чтобы получить всестороннюю характеристику разнообразия популяции [36]. Например, если разнообразие порядка $q = 0$ намного больше, чем разнообразие порядка $q = 1$ и 2, то это указывает на то, что в популяции имеется несколько редких аллелей, и частоты аллелей очень неравномерны. Также, очень похожие показатели разнообразия порядка $q = 1$ и 2 указывают на то, что в популяции доминируют несколько аллелей.

Непрерывный профиль разнообразия, который отображают числа Хилла порядка $q \geq 0$, содержит всю информацию о распределении аллелей (степени доминирования/концентрации) в популяции. Популяция с равночастотными аллелями будет иметь горизонтальный профиль разнообразия без какого-либо снижения. Популяция с высокой степенью концентрации аллелей будет иметь профиль разнообразия, резко снижающийся по мере увеличения q. Профили разнообразия – это наиболее полный и наглядный способ передачи информации о разнообразии систем (пород, видов, сортов). Их использование может способствовать принятию выверенных решений, например, по сохранению генофондных стад.

Заключение. Индекс Шеннона (${}^SH = -\sum p_m \ln p_m$) – это показатель «количества информации или возможности выбора, или энтропии», но не разнообразия. Экспоненциальная функция конвертирует SH в интуитивно понятный показатель разнообразия, выражаемый *эффективным числом элементов* (D).

В частности, показатель по объединённым субпопуляциям, S_{HGT} , конвертируется в D_γ -разнообразие – эффективное число аллелей по энтропии в популяции; средневзвешанный по субпопуляциям – S_{HWP} , в D_α -разнообразии – эффективное число аллелей по энтропии в усреднённой субпопуляции; межсубпопуляционная компонента – S_{HAP} , в D_β -разнообразии – эффективное число субпопуляций. Компонентам информационной меры Шеннона свойственна аддитивность ($S_{HGT} = S_{HWP} + S_{HAP}$), меры разнообразия Шеннона – мультипликативность ($D_\gamma = D_\alpha \times D_\beta$).

По семи породным выборкам абсолютное γ -разнообразие Шеннона составило 5,45, α -разнообразие – 3,76 *эффективных аллелей* [по энтропии], β -разнообразие – 1,45 *эффективных выборов*. Соответствующие масштабированные оценки (D' с диапазоном [0,1]) были 0,821, 0,765 и 0,362 (36,2 %). Последняя – сводная относительная оценка аллельной дифференциации породных выборов. Её дополнение, относительное перекрытие аллельных частот выборов, составило 63,8 %. Влияния уровня внутривыборочной гетерозиготности на полокусные S_{HAP} -оценки не было установлено (в отличие от $G_{ST(NEI)}$ и $F_{ST(W\&C)}$ мер), но имела место положительная тенденция D'_β -оценок.

Сводная по породным выборкам D'_β -оценка была в 3 раза выше оценок по $G_{ST(NEI)}$ и $F_{ST(W\&C)}$ мерам. Но последние не показатели аллельной дифференциации, а индексы фиксации. Они корректно указывали на то, что аллели в породных выборках очень далеки до полной гомозиготности. Индексы фиксации, скорректированные на уровень внутривыборочной гетерозиготности ($G''_{ST(HEd)}$ и $F'_{ST(W\&C)}$), производили оценки, которые были очень близки к D'_β . Несколько ниже, чем D'_β , но с 95 %-м доверительным интервалом $24 \div 34$ % были оценки по D_{JOST} и D_{CHAO} мерам.

Тест Мантеля показал высокую связь матрицы парных генетических дистанций (D'_β) между породами с таковыми, полученными по $G_{ST(NEI)}$, $F_{ST(W\&C)}$, $G''_{ST(HEd)}$, $F'_{ST(W\&C)}$, D_{JOST} и D_{CHAO} мерам ($R_M = 0,94 \div 0,98$). Как и в предыдущих исследованиях, на 2D-проекции D'_β -матрицы выделились группа «красных»

пород, группа голштинских «экотипов» и отдельно джерсейская порода. Ординация породных выборов имела достаточно большое сходство с ординациями по другим шести мерам разнообразия.

РСоА матрицы попарных оценок *несходства семи метрик* популяционной структуры визуализировал их ординацию на двумерной плоскости. В частности, выделялась мера аллельной дифференциации Шеннона (D'_β) и на значительном удалении от неё – семейство индексов фиксации ($G_{ST(NEI)}$ и $F_{ST(W\&C)}$) и семейство мер аллельной дифференциации на базе гетерозиготности ($G''_{ST(HEd)}$ и $F'_{ST(W\&C)}$, D_{JOST} и D_{CHAO}). Чтобы решить какую меру разнообразия/дифференциации следует использовать, надо знать, во-первых, что собираемся измерять, во-вторых, к чему чувствительна та или иная мера. Так, если задача исследования заключается в изучении демографических изменений и эволюционных историй популяций, то используются индексы фиксации ($G_{ST(NEI)}$ и $F_{ST(W\&C)}$). Если же ставится задача оценки актуальных генетических различий между (суб)популяциями, то лучше применять меры аллельной дифференциации ($G''_{ST(HEd)}$, $F'_{ST(W\&C)}$, D_{JOST} , D_{CHAO} , D'_β). При этом если интерес представляют только доминирующие (высокочастотные) аллели, то следует воспользоваться мерами $G''_{ST(HEd)}$, $F'_{ST(W\&C)}$, D_{JOST} и D_{CHAO} . Если же стоит задача получения наиболее полной информации об имеющихся аллельных различиях (суб)популяций, то лучшим является чувствительный шенноновский анализ (мера D'_β). Он взвешивает количество энтропии в m -ой аллели ($-\log p_m$) пропорционально её распространённости (p_m) и одновременно принимает во внимание несбалансированность (суб)популяций. Важно выбрать меру, которая будет отражать «эффект интереса», а затем интерпретировать фактические величины и их доверительные интервалы. Шенноновские информационно-энтропийные конструкции могут быть альтернативным подходом к количественной оценке биоразнообразия в иерархии от генов до экологических систем (включая различные виды генетической информации).

References

1. Хедрик Ф. Генетика популяций. М.: Техносфера, 2003. 592 с.
Hedrick P. W. *Genetika populyatsii* [Genetics of populations]. Moscow: «Technosfera», 2003. 592 p.
2. Nei M. Analysis of gene diversity in subdivided populations. Proc. Nat. Acad. Sci. USA. 1973;70(12):3321-3323.
DOI: <https://doi.org/10.1073/pnas.70.12.3321>
3. Jost L. Partitioning diversity into independent alpha and beta components. Ecology. 2007;88(10):2427-2439.
DOI: <https://doi.org/10.1890/06-1736.1>

4. Jost L., Archer F., Flanagan S., Gaggiotti O., Hoban S., Latch E. Differentiation measures for conservation genetics. *Evol. Appl.* 2018;11(7):1139-1148. DOI: <https://doi.org/10.1111/eva.12590>
5. Weir B. S., Cockerham C. C. Estimating F-statistics for the analysis of population structure. *Evolution*. 1984;38(6):1358-1370. DOI: <https://doi.org/10.2307/2408641>
6. Meirmans P. G., Hedrick P. W. Assessing population structure: F_{ST} and related measures. *Mol. Ecol. Res.* 2011;11(1):5-18. DOI: <https://doi.org/10.1111/j.1755-0998.2010.02927.x>
7. Jost L. G_{ST} and its relatives do not measure differentiation. *Mol. Ecol.* 2008;17(18):4015-4026. DOI: <https://doi.org/10.1111/j.1365-294X.2008.03887.x>
8. Кузнецов В. М. F-статистики Райта: оценка и интерпретация. Проблемы биологии продуктивных животных. 2014;(4):80-104. Режим доступа: <https://www.elibrary.ru/item.asp?id=22833217>
Kuznetsov V. M. *F-statistiki Rayta: otsenka i interpretatsiya*. [Wright's F-statistics: estimation and interpretation]. *Problemy biologii produktivnykh zhivotnykh* = Problems of Productive Animal Biology. 2014;(4):80-104. (In Russ.). URL: <https://www.elibrary.ru/item.asp?id=22833217>
9. Кузнецов В. М. Методы Ней для анализа генетических различий между популяциями. Проблемы биологии продуктивных животных. 2020;(1):91-110. DOI: <https://doi.org/10.25687/1996-6733.prodanimbiol.2020.1.91-110>
Kuznetsov V. M. *Metody Ney dlya analiza geneticheskikh razlichiy mezhdu populyatsiyami*. [Nei's methods for analyzing genetic differences between populations]. *Problemy biologii produktivnykh zhivotnykh* = Problems of Productive Animal Biology. 2020;(1):91-110. (In Russ.). DOI: <https://doi.org/10.25687/1996-6733.prodanimbiol.2020.1.91-110>
10. Кузнецов В. М. Сравнение методов оценки генетической дифференциации популяций по микросателлитным маркерам. Аграрная наука Евро-Северо-Востока. 2020;21(2):169-182. DOI: <https://doi.org/10.30766/2072-9081.2020.21.2.169-182>
Kuznetsov V. M. *Sravnenie metodov otsenki geneticheskoy differentsiatsii populyatsiy po mikrosatelitnym markeram*. [Comparison of methods for evaluating genetic differentiation of populations by microsatellite markers]. *Agrarnaya nauka Evro-Severo-Vostoka* = Agricultural Science Euro-North-East. 2020;21(2):169-182. (In Russ.). DOI: <https://doi.org/10.30766/2072-9081.2020.21.2.169-182>
11. Кузнецов В. М. Оценка генетической дифференциации популяций молекулярным дисперсионным анализом (аналитический обзор). Аграрная наука Евро-Северо-Востока. 2021;22(2):167-187. DOI: <https://doi.org/10.30766/2072-9081.2021.22.2.167-187>
Kuznetsov V. M. *Otsenka geneticheskoy differentsiatsii populyatsiy molekulyarnym dispersionnym analizom (analiticheskiy obzor)*. [Assessment of genetic differentiation of populations by analysis of molecular variance (analytical review)]. *Agrarnaya nauka Evro-Severo-Vostoka* = Agricultural Science Euro-North-East. 2021;22(2):167-187. (In Russ.). DOI: <https://doi.org/10.30766/2072-9081.2021.22.2.167-187>
12. Lewontin R. C. The apportionment of human diversity. *J. Evol. Biol.* 1972;6:381-398. DOI: https://doi.org/10.1007/978-1-4684-9063-3_14
13. Hill M. O. Diversity and evenness: A unifying notation and its consequence. *Ecology*. 1973;54(2):427-432. DOI: <https://doi.org/10.2307/1934352>
14. Shannon C. E. A Mathematical Theory of Communication. Reprinted with corrections from. *The Bell System Technical Journal*. 1948;27(3):379-423, 623-656. URL: <https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf>
15. Ashby W. R. An introduction to cybernetics. Chapman & Hall LTD, London, 1956. 295 p. URL: <https://archive.org/details/introductiontocy00ashb/page/294/mode/2up>
16. Corso G., Ferreira G. M. F., Lewinsohn T. M. Mutual Information as a General Measure of Structure in Interaction Networks. *Entropy* 2020;22(5): 528. DOI: <https://doi.org/10.3390/e22050528>
17. Sheldon A. L. Equitability indices: Dependence on the species count. *Ecology*. 1969;50(3): 466-467. DOI: <https://doi.org/10.2307/1933900>
18. Sherwin W. B. Entropy and Information Approaches to Genetic Diversity and its Expression: Genomic Geography. *Entropy*. 2010;12(7):1765-1798. DOI: <https://doi.org/10.3390/e12071765>
19. Харзинова В. Р., Гладырь Е. А., Федоров В. И., Романенко Т. М., Шимит Л. Д., Лайшев К. А., Калашникова Л. А., Зиновьева Н. А. Разработка мультиплексной панели микросателлитов для оценки достоверности происхождения и степени дифференциации популяций северного оленя *Rangifer Tarandus*. Сельскохозяйственная биология. 2015;50(6):756-765. DOI: <https://doi.org/10.15389/agrobiology.2015.6.756rus>
Kharzinova V. R., Gladys' E. A., Fedorov V. I., Romanenko T. M., Shimit L. D., Layshev K. A., Kalashnikova L. A., Zinovieva N. A. *Raz-rabotka mul'tipleksnoy paneli mikrosatelлитov dlya otsenki dostovernosti proiskhozhdeniya i stepeni differentsiatsii populyatsiy severnogo olenya Rangifer Tarandus*. [Development of multiplex microsatellite panel to assess the parentage verification in and differentiation degree of reindeer populations (*Rangifer Tarandus*)]. *Sel'skokhozyaystvennaya biologiya* = Agricultural Biology. 2015;50(6): 756-765. (In Russ.). DOI: <https://doi.org/10.15389/agrobiology.2015.6.756rus>
20. Денискова Т. Е., Гладырь Е. А., Зиновьева Н. А. Характеристика некоторых российских пород овец по микросателлитным маркерам. Актуальные проблемы гуманитарных и естественных наук. 2016;(9-1):24-29. Режим доступа: <https://www.elibrary.ru/item.asp?id=26704429>
Deniskova T. E., Gladys' E. A., Zinovieva N. A. *Kharakteristika nekotorykh rossiyskikh porod ovets po mikrosatelitnym markeram*. [Characteristics of some Russian sheep breeds by microsatellite markers]. *Aktual'nye problemy gumanitarnykh i estestvennykh nauk*. 2016;(9-1): 24-29. (In Russ.). URL: <https://www.elibrary.ru/item.asp?id=26704429>
21. Rossetto M., Kooyman R., Sherwin W. B., Jones R. Dispersal limitations, rather than bottlenecks or habitat specificity, can restrict the distribution of rare and endemic rainforest trees. *Amer. J Botany*. 2008;95(3):321-329. DOI: <https://doi.org/10.3732/ajb.95.3.321>

22. Peakall R., Smouse P. E. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology* 2006;6(1):288-295. DOI: <https://doi.org/10.1111/j.1471-8286.2005.01155.x>
23. Peakall R., Smouse P. E. GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics*. 2012;28(19):2537-2539. DOI: <https://doi.org/10.1093/bioinformatics/bts460>
24. Smouse P. E., Whitehead M., Peakall R. An informational diversity framework, illustrated with sexually deceptive orchids in early stages of speciation. *Mol. Ecol. Resour.* 2015;15(6):1375-1384. DOI: <https://doi.org/10.1111/1755-0998.12422>
25. STATGRAPHICS® Centurion XVI User Manual. By StatPoint Technologies, Inc. 2010. 297 p.
26. Kimura M., Crow J. F. The number of alleles that can be maintained in a finite population. *Genetics*. 1964;49(4):725-738. DOI: <https://doi.org/10.1093/genetics/49.4.725>
27. Jost L. Entropy and diversity. *Oikos*. 2006;113(2):363-375. DOI: <https://doi.org/10.1111/j.2006.0030-1299.14714.x>
28. Sherwin W. B., Jabot F., Rush R., Rossetto M. Measurement of biological information with applications from genes to landscapes. *Molec. Ecol.* 2006;15(10):2857-2869. DOI: <https://doi.org/10.1111/j.1365-294x.2006.02992.x>
29. Chao A., Ma K. H., Hsieh T. C., Chiu C. H. Online Program SpadeR (Species-richness Prediction And Diversity Estimation in R). Program and User's Guide. 2015. URL: http://chao.stat.nthu.edu.tw/wordpress/software_download/
30. Ma L., Ji Y.-J., Zhang D.-X. Statistical measures of genetic differentiation of populations: Rationales, history and current states. *Current Zoology*. 2015;61(5):886-897. DOI: <https://doi.org/10.1093/czoolo/61.5.886>
31. Доцев А. В., Сермягин А. А., Шахин А. В., Паронян И. А., Племяшов К. В., Рейер Х., Виммерс К., Брем Г., Зиновьева Н. А. Оценка современного состояния генофонда холмогорской и чернопестрой пород крупного рогатого скота на основе полногеномного SNP-анализа. *Вавиловский журнал генетики и селекции*. 2018;22(6):742-747. DOI: <https://doi.org/10.18699/VJ18.418>
- Dotsev A. V., Sermyagin A. A., Shakhin A. V., Paronyan I. A., Plemyashov K. V., Reyer H., Wimmers K., Brem G., Zinovieva N. A. *Otsenka sovremennogo sostoyaniya genofonda kholmogorskoj i cherno-pestroy porod krupnogo rogatogo skota na osnove polnogenomnogo SNP-analiza*. [Evaluation of current gene pool of Kholmogor and Black-and-white cattle breeds based on whole genome SNP analysis]. *Vavilovskiy zhurnal genetiki i selektsii* = Vavilov Journal of Genetics and Breeding. 2018;22(6): 742-747. (In Russ.). DOI: <https://doi.org/10.18699/VJ18.418>
32. Волкова В. В., Романенкова О. С., Денискова Т. Е., Мишина А. И., Костюнина О. В., Зиновьева Н. А. Характеристика аллелофонда холмогорской породы крупного рогатого скота с использованием STR-маркеров. *Молочное и мясное скотоводство*. 2019;(7): 3-7. Режим доступа: <https://elibrary.ru/item.asp?id=41852279>
- Volkova V. V., Romanenkova O. S., Deniskova T. E., Mishina A. I., Kostyunina O. V., Zinovieva N. A. *Kharakteristika allelofonda kholmogorskoj porody krupnogo rogatogo skota s ispol'zovaniem STR-markerov*. [Assessment of the allele pool of the kholmogory cattle breed with using STR-markers]. *Molochnoe i myasnoe skotovodstvo* = Journal of Dairy and Beef Cattle Farming. 2019;(7):3-7. (In Russ.). URL: <https://elibrary.ru/item.asp?id=41852>
33. Abdelmanova A. S., Kharzinova V. R., Volkova V. V., Mishina A. I., Dotsev A. V., Sermyagin A. A., Boronetskaya O. I., Petrikeeva L. V., Chinarov R. Yu., Brem G., Zinovieva N. A. Genetic diversity of the historical and modern populations of Russian cattle breeds revealed by microsatellite analysis. *Genes*. 2020;11(8):940. DOI: <https://doi.org/10.3390/genes11080940>
34. Трипольский К. С. Курс скотоводства. Руководство к разведению рогатого скота, овец, лошадей и свиней. Часть 2. Частное скотоводство. С.-Петербург: Издание А. Ф. Девриена, 1875. 429 с.
- Tripol'skiy K. S. *Kurs skotovodstva. Rukovodstvo k razvedeniyu rogatogo skota, ovets, loshadey i sviney*. [Cattle breeding course. A guide to breeding cattle, sheep, horses and pigs]. Part. 2. *Chastnoe skotovodstvo*. [Private cattle breeding]. S.-Petersburg: *Izdanie A. F. Devriena*, 1875. 429 p.
35. Ricotta C. On parametric evenness measures. *J. of Theoret. Biol.* 2003;222(2):189-197. DOI: [https://doi.org/10.1016/S0022-5193\(03\)00026-2](https://doi.org/10.1016/S0022-5193(03)00026-2)
36. Gaggiotti O. E., Chao A., Peres-Neto P., Chiu C.-H., Edwards C., Fortin M.-J., Jost L., Richards C. M., Selkoe K. A. Diversity from genes to ecosystems: A unifying framework to study variation across biological metrics and scales. *Evol Appl.* 2018;11(7):1176-1193. DOI: <https://doi.org/10.1111/eva.12593>

Сведения об авторе

✉ Кузнецов Василий Михайлович, доктор с.-х. наук, профессор, зав. лабораторией популяционной генетики в животноводстве, ФГБНУ «Федеральный аграрный научный центр Северо-Востока имени Н. В. Рудницкого», ул. Ленина, д. 166а, г. Киров, Российская Федерация, 610007, e-mail: priemnaya@fanc-sv.ru, ORCID: <https://orcid.org/0000-0002-2219-805X>, e-mail: vm-kuznetsov@mail.ru

Information about the author

✉ Vasilii M. Kuznetsov, DSc in Agricultural Science, professor, Head of the Laboratory of Population Genetics in Animal Husbandry, Federal Agricultural Research Center of the North-East named N. V. Rudnitsky, Lenin str., 166a, Kirov, Russian Federation, 610007, e-mail: priemnaya@fanc-sv.ru, ORCID: <https://orcid.org/0000-0002-2219-805X>, e-mail: vm-kuznetsov@mail.ru

✉ – Для контактов / Corresponding author