https://doi.org/10.30766/2072-9081.2023.24.5.888-906
УДК 636.082:575.174.015.3

Анализ полокусных оценок аллельного разнообразия STR-маркеров в выборке быков-производителей

© 2023. В. М. Кузнецов ✉
ФГБНУ «Федеральный аграрный научный центр Северо-Востока имени Н. В. Рудницкого», г. Киров, Российская Федерация

Генотипы по 11 локусам микросателлитов ДНК 84 быков семи пород использовали для полокусной оценки 14 показателей аллельного разнообразия/дифференциации. К сформированным матрицам данных из оригинальных и преобразованных оценок размерностью 11×14 применены традиционные и многомерные методы статистики. Наименьшей изменчивостью характеризовались оценки гетерозиготности – 8-14 %. Изменчивость числа аллелей на локус и показателей дифференциации пород была на уровне 20-26%, индексов фиксации – 38-44 %. Установлены статистически значимые корреляции Кендалла (0,8-1,0) между показателями аллельного богатства и гетерозиготности, индексами фиксации, показателями дифференциации. Изменчивость преобразованных оценок показателей разнообразия/дифференциации в пределах локуса была в диапазоне 6-32%, в том числе, по локусам Eth3, Tgla122, Eth225, Bm2113 – 6-12%, локусам Inra23, Tgla126, Eth10 – 15-20 %, локусам Tgla227, Sps115, Tgla53, Bm1824 – 28-32 %. Непараметрический тест Mann-Whitney-Wilcoxon показал статистически значимые различия медиан локуса Eth3 с локусом Bm2113, локуса Tgla126 с локусами Eth3, Inra23, Tgla122, Eth225, Bm2113, Bm1824, Eth10. Анализ главных компонент (PCA) выделил две компоненты с общей информативностью 95,2 %. Первая учитывала 59,4 % общей дисперсии, имела наибольшие нагрузки по показателям внутрипородного разнообразия и была определена как «альфа-компонента». Вторая объясняла 35,8 % общей дисперсии, имела высокие нагрузки по показателям межпородной дифференциации и была определена как «бета-компонента». 2D-PCA-ординация показала, что для анализируемых породных выборок, локусов и мер разнообразия имела место характерная группировка локусов. Локусы Tgla227 и Tgla53 сформировали группу A, группу B – локусы Tgla122, Eth225, Eth10, группу C – локусы Inra23, Bm2113 и Bm1824. Локусы условной группы D (Eth3, Tgla126, Sps115) были определены как «нетипичные». Валидность ординации подтверждали расчётами по редуцированным данным (размерностью 11×7) и методом неметрического многомерного шкалирования (nMDS). Согласованность ординаций по тесту Прокруста была 96 % (p_{perm} = 0,001). Аналогичную классификацию локусов дал кластерный анализ (UPGMA) с бутстрэп-вероятностями кластера A – 73 %, B – 100 %, C – 73 %, D – 47 %. Были рассчитаны дистанции и показатели сходства (S) профилей локусов со сводными оценками по 11 локусам (определены как «истинные»). Локусы Tgla126 и Sps115 имели S ≈ 40 %, Tgla53 и Bm1824 – на уровне 60 %; Inra23, Tgla227 и Bm2113 – 70-75 %, локусы Eth3, Tgla122, Eth225 и Eth10 – 84-88 %. Среднее абсолютное отклонение оценок показателей разнообразия по четырём локусам с S≥84 % от «истинных» оценок было 3,4 %, по четырём локусам с S≤60 % – 12,4 %. По компонентным оценкам для каждого локуса был рассчитан тотальный показатель разнообразия (γLV). Линейная связь γLV с полокусными оценками γ-разнообразия с вероятностью 95 % находилась в интервале 0,73-0,98, ранговая корреляция Кендалла была 0,67 (p_{value} = 0,005). Проведённое исследование вносит определённый вклад в расширение инструментариев для обработки молекулярно-генетических данных при анализе аллельного разнообразия в подразделённых популяциях.

Ключевые слова: *крупный рогатый скот, локусы, микросателлиты, разнообразие, дифференциация, методы многомерной статистики, 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 123011900029-6).

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

***Конфликт интересов**:* автор заявил об отсутствии конфликта интересов.

***Для цитирования*:** Кузнецов В. М. Анализ полокусных оценок аллельного разнообразия STR-маркеров в выборке быков-производителей. Аграрная наука Евро-Северо-Востока. 2023;24(5):888-906.
DOI: https://doi.org/10.30766/2072-9081.2023.24.5.888-906

Поступила: 17.02.2023 Принята к публикации: 31.08.2023 Опубликована онлайн: 30.10.2023

Analysis of locus estimates of allelic diversity of STR markers in a sample of breeding bulls

© 2023. Vasiliy M. Kuznetsov ✉
Federal Agricultural Research Center of the North-East named N. V. Rudnitsky, Kirov, Russian Federation

Genotypes of the 11 DNA microsatellite loci of 84 bulls of seven breeds were used to evaluate 14 indicators of allelic diversity/differentiation. Traditional and multidimensional statistical methods were applied to the data matrices from the original and transformed estimates (11×14). Estimates of heterozygosity had coefficients of variability of 8-14 %, the number

of alleles per locus and indicators of differentiation of breeds at the level of 20-26 %, fixation indices – 38-44 %. Statistically significant Kendall correlations (0.8-1.0) between indicators of allelic richness and heterozygosity, fixation indices, and differentiation indicators were established. The variability of the transformed estimates of diversity/differentiation indicators by loci was in the range of 6-32 %. Including by loci Eth3, Tgla122, Eth225, Bm2113 – 6-12 %, loci Inra23, Tgla126, Eth10 – 15-20 %, loci Tgla227, Sps115, Tgla53, Bm1824 – 28-32 %. The nonparametric Mann-Whitney-Wilcoxon test showed statistically significant differences in the medians of the Eth3 locus with the Bm2113 locus, the Tgla126 locus with the Eth3, Inra23, Tgla122, Eth225, Bm2113, Bm1824, Eth10 loci. The principal component analysis (PCA) identified two components with a total information content of 95,2 %. The first one took into account 59.4 % of the total variance, had the highest loads in intra-breed diversity data and was defined as an «alpha component». The second accounted for 35.8 % of the total variance, had the highest loads in inter-breed differentiation data and was defined as a «beta component». 2D-PCA-ordination showed that a characteristic grouping of loci took place for the analyzed breeds (samples), loci and measures of diversity. Loci Tgla227 and Tgla53 formed group A, group B – loci Tgla122, Eth225, Eth10, group C – loci Inra23, Bm2113 and Bm1824. The loci of the conditional group D (Eth3, Tgla126, Sps115) were defined as «untypical». Validation of ordination was confirmed by calculations on reduced data (dimension 11×7) and the method of non-metric multidimensional scaling (nMDS). The consistency of ordinations according to the Procrust test was 96 % (p_{perm} <0.001). A similar classification of loci was obtained by cluster analysis (UPGMA) with butstrap probabilities of cluster: A – 73, B – 100, C – 73, D – 47 %. The distances and similarity indicators (S) between the profiles of loci and the «true» summary estimates for 11 loci were calculated. Loci Tgla126 and Sps115 had S ≈ 40 %, loci Tgla53 and Bm1824 – at the level of 60 %, loci Inra23, Tgla227 and Bm2113 – 70-75 %, loci Eth3, Tgla122, Eth225 and Eth10 – 84-88 %. The average absolute deviation of the estimates of diversity indicators for the four loci with S≥84 % from the «true» estimates was 3.4 %, for the four loci with S≤60 % – 12.4 %. According to component scores, a general diversity index, γLV, was calculated for each locus. Its correlation with the estimates of the Shannon/Sherwin's γ-diversity with a 95 % probability value was in the range of 0.73-0.98, Kendall's rank correlation was 0.67 (p_{value} = 0.005). The conducted research makes a certain contribution to the expansion of tools for processing molecular genetic data in the analysis of allelic diversity in subdivided populations.

Keywords: *cattle, loci, microsatellites, diversity, differentiation, methods of multidimensional statistics, 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. 123011900029-6).

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

Conflict of Interest: the author declared no conflicts of interest.

For citation: Kuznetsov V. M. Analysis of locus estimates of allelic diversity of STR markers in a sample of breeding bulls. *Agrarnaya nauka Evro-Severo-Vostoka* = Agricultural Science Euro-North-East. 2023;24(5):888-906. (In Russ.). DOI: https://doi.org/10.30766/2072-9081.2023.24.5.888-906

Received: 17.02.2023 Accepted for publication: 31.08.2023 Published online: 30.10.2023

В последние годы ДНК-маркеры находят всё большее применение в исследованиях по генетике, селекции, разведению коммерческих и сохранению вытесняемых российских пород сельскохозяйственных животных. Изучается эффективность использования микросателлитов и однонуклеотидного полиморфизма [1], рассчитывается геномная племенная ценность [2] и геномный инбридинг животных [3], оценивается генетическое разнообразие внутри (α) и между (β) породами крупного рогатого скота [4, 5], лошадей [6], овец [7], свиней [8], оленей [9].

Для измерения α- и β-разнообразия был предложен ряд статистик: Райта, Нея, Вейра и Кокерхэма, Джоста, Чао, Шеннона/Шервина [10, 11, 12, 13, 14, 15], которые базируются на разных биологических и математических допущениях. Проводится сравнительный анализ этих мер разнообразия [16, 17, 18]. Все они способствуют получению той или иной информации о различных аспектах аллельного разнообразия, демографической истории и/или дифференциации пород, о структуре генетической изменчивости популяций.

В исследованиях подобного рода обычно анализируются несколько породных выборок и ряд локусов, по которым рассчитываются сводные оценки разнообразия/дифференциации и/или парные (между породами) генетические дистанции. Как представляется, не менее интересным и полезным может быть сравнительный анализ оценок разнообразия/дифференциации по каждому локусу и поиск путей использования разных показателей для комплексной характеристики генетической изменчивости. Например, векторы/профили полокусных оценок побуждают к использованию методов ординации, которые позволяют обнаружить скрытые (латентные) обобщающие характеристики организационной структуры, визуализировать сходство и различие изучаемых локусов, их кластеризацию. Или, если в рамках нейтральной модели доверительные интервалы оценок показателей разнообразия (например,

по F_{ST}) по каким-либо локусам не перекрываются ($p_{value} \leq 0{,}05$), то это может указывать на наличие возмущающего фактора(-ов), что, в свою очередь, может мотивировать проведение дальнейших углублённых исследований.

Цель настоящей работы – полокусная оценка разными методами α- и β-разнообразия в выборке животных, генотипированных по ДНК-маркерам, с последующим традиционным и многомерным статистическим анализами полученных показателей для выявления неслучайных межлокусных отношений.

Материал и методы. В работе использовали те же данные, что и в предыдущих публикациях [16, 17, 18]. В частности, 84 быка, каждый генотипирован по 11 STR-локусам (микросателлиты ДНК)[1]: 10 быков джерсейской породы, 10 – айрширской породы, 10 – красной датской, 9 – красной шведской и 45 быков голштинской породы трёх «экотипов» (отродий): 13 быков из Германии, 17 – из Нидерландов, 15 – из США.

По каждому локусу рассчитывали 14 показателей (характеристик) аллельного разнообразия:

- «аллельное богатство» на локус (n_a – фактическое число аллелей, n_e – число эффективных аллелей по гетерозиготности, $^{S}n_e$ – число эффективных аллелей по энтропии);
- гетерозиготность (H_o – фактическая, H_e – ожидаемая);
- индексы фиксации ($G_{ST(NEI)}$ – по Нею, $F_{ST(W\&C)}$ – по Кокерхэм и Вейру);
- дифференциация породных выборок на базе гетерозиготности ($G''_{ST(HED)}$ и $F'_{ST(W\&C)}$ – модифицированные индексы фиксации, D_{JOST} – по Джосту, D_{CHAO} – по Чао) и
- α-, β- и γ-разнообразие на базе энтропии по Шеннону/Шервину (D'_{α} – внутривыборочное, D'_{β} – межвыборочное, D'_{γ} – общее).

Показатели первых двух пунктов относились к α-разнообразию, следующих двух – к β-разнообразию, которые, при наличии нескольких породных выборок, характеризовали изменчивость оценок α-разнообразия в пространстве. Методические стороны оценивания рассмотрены в [10, 11, 16, 17, 18]. Для анализа полокусных оценок применяли методы традиционной (описательная статистика, парное сравнение средних и медиан, ковариационный анализ) и многомерной статистики (анализ главных компонент, неметрическое многомерное шкалирование, кластерный анализ и анализ Прокруста). Использовали компьютерные программы GenAlEx 6.502 [20, 21, 22], PAST3 [23], KyPlot 6.0 [24], STATGRAPHICS® Centurion XVI[2], прокрустов анализ проводили по алгоритмам [25, 26, 27].

Основная часть. *Оценки показателей разнообразия*. В таблице 1 даны полокусные оценки показателей аллельного разнообразия/дифференциации; в последнем столбце – сводные оценки по всем локусам[3].

Каждый из 11 локусов был представлен вектором оценок 14 характеристик-признаков – показателей разнообразия, которые формировали набор данных размерностью 11×14 (**Data1**). Число аллелей на локус (n_a) было в диапазоне от 3,9 (локус 6) до 7 (локус 8), эффективных аллелей по энтропии ($^{S}n_e$) – от 2,5 (локус 6) до 5,4 (локус 3), наблюдаемой гетерозиготности (H_o) – от 49 до 78 % (локусы 6 и 8), индекса фиксации по Нею ($G_{ST(NEI)}$) – от 4,9 до 18,6 % (локусы 8 и 10), дифференциации породных выборок по Шеннону/Шервину – от 22,1 до 43,8 % (локусы 4 и 2) и т.п. Все переменные «прошли» W-тест (Shapiro-Wilk) на нормальность распределения. Коэффициенты изменчивости (CV) оценок D'_{α} и D'_{γ} составили 8-10 %; H_o и H_e – 14 %; n_a, n_e, $^{S}n_e$, $G''_{ST(HED)}$, $F'_{ST(W\&C)}$, D_{JOST}, D_{CHAO} и D'_{β} – 20-26 %; $G_{ST(NEI)}$, $F_{ST(W\&C)}$ (индексы фиксации) – 38-44 %.

Величины переменных таблицы 1 (p_{km}) имели разные шкалы и единицы измерения, что делало невозможным применение статистики по локусам. Поэтому p_{km} были преобразованы к единому масштабу[4]. «Обезразмеренную» переменную (p'_{km}) можно интерпретировать как «процент вклада k-го локуса в сводную оценку m-ой переменной по всем локусам».

[1]STR – Simple Tandem Repeats – участки ДНК длиной 2-6-9 нуклеотидов, тандемно повторенных 5-40 раз с частотой мутаций (изменений в геноме) от 10^{-6} до 10^{-2} в разных локусах (10^{-6} – одна мутация на один миллион событий репликации – удвоения молекулы ДНК). Аллели с более высоким числом повторов часто мутируют с более высокой скоростью; связь экспоненциальная [19].

[2]STATGRAPHICS® Centurion XVI User Manual. By StatPoint Technologies, Inc. 2010. 297 p.

[3]Из 77 тестов согласия распределений генотипов с равновесием Харди-Вайнберга (РХВ) в двух случаях имело место статистически значимое отклонение от РХВ: по локусу Eth10 в выборке быков джерсейской породы и по локусу Eth225 в выборке голштинских быков из Германии.

[4]Использовалась формула: $p'_{km}=100\times p_{km}/(p_{1m}.+ p_{2m}...+...p_{km}...+...p_{Km})$, где k – локус, K – число локусов, m – переменная.

Таблица 1 – **Оценки показателей разнообразия 11 STR-локусов (Data1:** 11×14) /
Table 1 – **Estimates of diversity indicators of 11 STR loci (Data1:** 11×14)

Variable	*Eth3*	*Inra23*	*Tgla227*	*Tgla126*	*Tgla122*	*Sps115*	*Eth225*	*Tgla53*	*Bm2113*	*Bm1824*	*Eth10*	*1-11*
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*	*(10)*	*(11)*	
n_a*	4,1	4,1	6,9	4,1	5,7	3,9	5,4	7,0	5,0	3,9	5,3	5,0
n_e	2,7	2,9	4,7	2,4	3,9	2,1	3,7	4,4	3,8	2,5	4,1	3,4
^Sn_e*	3,0	3,2	5,4	2,9	4,4	2,5	4,1	5,1	3,9	2,8	4,4	3,7
H_o, %	79	76	79	60	79	54	81	84	84	61	78	74
H_e, %*	61	62	78	58	72	49	71	75	69	56	75	66
$G_{ST(NEI)}$, %*	11,8	14,6	5,0	7,4	8,2	16,7	9,4	4,9	12,1	18,6	7,2	10,3
$F_{ST(W\&C)}$, %	12,8	14,8	5,7	7,8	9,3	16,0	10,8	5,8	12,2	18,3	8,0	10,8
$G''_{ST(Hed)}$, %	36,2	47,1	31,8	21,7	37,6	38,5	40,8	26,6	47,6	50,5	37,4	37,5
$F'_{ST(W\&C)}$, %*	34,3	43,9	31,2	19,5	39,0	33,8	39,1	28,0	43,9	45,9	36,3	35,1
D_{JOST}, %	26,2	36,5	27,6	14,4	31,1	24,2	33,6	22,2	39,3	37,4	31,7	29,2
D_{CHAO}, %*	24,7	35,8	28,1	14,4	30,7	24,0	32,5	21,4	38,1	37,3	31,7	30,8
D'_α, %	70,1	73,0	85,4	67,0	81,4	64,9	77,7	84,8	78,5	68,3	81,2	76,5
D'_β, %*	33,3	43,8	38,1	22,1	40,1	22,5	39,2	41,8	38,4	34,0	42,0	36,2
D'_γ, %	77,0	81,7	88,4	71,4	86,1	69,9	83,6	88,5	83,9	76,6	86,3	82,1

Примечания. Число аллелей на локус: n_a – фактическое, n_e – эффективное по гетерозиготности, ^Sn_e – эффективное по энтропии. Гетерозиготность: H_o – фактическая, H_e – ожидаемая. Индекс фиксации: $G_{ST(NEI)}$ – по Нею, $F_{ST(W\&C)}$ – по Кокерхэм и Вейру. Показатели дифференциации пород на базе гетерозиготности: $G''_{ST(HED)}$ и $F'_{ST(W\&C)}$ – скорректированные индексы фиксации, D_{JOST} – по Джосту, D_{CHAO} – по Чао. Энтропийные меры разнообразия по Шеннону/Шервину: D'_α – внутри пород, D'_β, – между породами, D'_γ – общее. * переменные, которые формировали Data1r размерностью 11×7 /

Notes. The number of alleles per locus: n_a – actual, n_e – effective in heterozygosity, ^Sn_e – effective in entropy. Heterozygosity: H_0 – actual, H_e – expected. Fixation index: $G_{ST(NEI)}$ – by Hei′s, $FST_{(W\&C)}$ – by Cockerham/Weir′s. Indicators of interbreed differentiation based on heterozygosity: $G''_{ST(HED)}$ and $F'_{ST(W\&C)}$ – adjusted fixation indices, D_{JOST} – by Jost′s, D_{CHAO} – by Chao′s. Entropic measures of diversity according to Shannon/Sherwin′s: D'_α –intrabreed, D'_β, – between breeds, D'_γ – in combined samples. * variables that formed Data1r with dimension 11×7.

Преобразованные оценки представлены в таблице 2 (**Data2**). Каждый столбец рассматривался как профиль локуса k по M = 14 преобразованным оценкам разнообразия/дифференциации, а каждая строка – как профиль переменной m по K = 11 полокусным оценкам.

Внутрилокусный CV p'_{km}-оценок варьировал от 6 до 32 %, в частности, по локусу 7 – 6,0 %, по локусам 1, 5 и 9 – 10-12 %, локусам 2, 4 и 11 – 15-20 %, локусам 3, 6, 8 и 10 – 28-32 %. Усреднённый CV составил 19,5 %, средних по локусам – 11,5 %.

На рисунке 1 представлены кривые (профили) полокусных оценок показателей разнообразия. Так, на рисунке 1, А визуализированы профили показателей аллельного богатства (n_a, n_e и ^Sn_e); кривые – почти идентичны. Высокие значения имели локусы 3 и 8 (в среднем 12,4 %), низкие – локусы 4 и 6 (6,6 %); CV оценок составил 20-26 %. Схожие кривые, но менее вариативные (CV = 14 %), получены для профилей фактической (H_o) и ожидаемой (H_e) гетерозиготности (рис. 1, В).

Кривые профилей индексов фиксации и показателей β-разнообразия на базе гетерозиготности представлены на рисунке 1, С. Выделялась кривая для индексов фиксации (идентичная для $G_{ST(NEI)}$ и $F_{ST(W\&C)}$) из-за высокой вариативности оценок (CV = 40 % против 24 %). Имела место некоторая асинхронность с кривыми α-разнообразия (рис. 1, А и 1, В). Так, если максимальные оценки по α-разнообразию получены у локусов 3, 5 и 8, то оценки их индексов фиксации были минимальными. С другой стороны, если по локусам 2, 6 и 10 индексы фиксации были самыми высокими, то оценки их α-разнообразия – самыми низкими. В более полиморфных локусах скорость фиксации аллелей ниже.

Таблица 2 – **Преобразованные оценки разнообразия, % (Data2: 11×14) /**
Table 2 – **Transformed diversity estimates, % (Data2: 11×14)**

Variable	*Локус / Locus*										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
n_a*	7,4	7,4	12,5	7,4	10,3	7,0	9,7	12,6	9,0	7,0	9,6
n_e	7,3	7,8	12,6	6,5	10,5	5,6	9,9	11,8	10,2	6,7	11,0
$^{S}n_e$*	7,2	7,7	12,9	7,0	10,6	6,0	9,8	12,2	9,4	6,7	10,6
H_o	9,7	9,3	9,7	7,4	9,7	6,6	9,9	10,3	10,3	7,5	9,6
H_e*	8,4	8,5	10,7	8,0	9,9	6,7	9,8	10,3	9,5	7,7	10,3
$G_{ST(NEI)}$*	10,2	12,6	4,3	6,4	7,1	14,4	8,1	4,2	10,4	16,0	6,2
$F_{ST(W\&C)}$	10,5	12,2	4,7	6,4	7,7	13,2	8,9	4,8	10,0	15,1	6,6
$G''_{ST(Hed)}$	8,7	11,3	7,6	5,2	9,0	9,3	9,8	6,4	11,4	12,1	9,0
$F'_{ST(W\&C)}$*	8,7	11,1	7,9	4,9	9,9	8,6	9,9	7,1	11,1	11,6	9,2
D_{Jost}	8,1	11,3	8,5	4,4	9,6	7,5	10,4	6,8	12,1	11,5	9,8
D_{Chao}*	7,8	11,2	8,8	4,5	9,6	7,5	10,2	6,7	12,0	11,7	9,9
D'_{α}	8,4	8,8	10,3	8,0	9,8	7,8	9,3	10,2	9,4	8,2	9,8
D'_{β}*	8,4	11,1	9,6	5,6	10,1	5,7	9,9	10,6	9,7	8,6	10,6
D'_{γ}	8,6	9,1	9,9	8,0	9,6	7,8	9,4	9,9	9,4	8,6	9,7
Mean	8,5	9,9	9,2	6,5	9,5	8,3	9,6	8,8	10,2	10,0	9,3
CV	11,9	17,8	28,3	20,3	10,5	32,4	6,0	31,4	9,8	31,0	14,8

Примечания: Mean – среднее по локусу; CV – коэффициент изменчивости. * переменные, которые формировали **Data2r** размерностью 11×7 (как в **Data1r**) /

Notes: Mean – the mean by locus; CV – the coefficient of variability. * variables that formed a **Data2r** with a dimension of 11×7 (as in **Data1r**).

На рисунке 1, D представлены три кривые, характеризующие вариацию полокусных оценок *α*-, *β*- и *γ*-разнообразия по энтропии. Профили D'_{α}- и D'_{γ}-переменных были сильно коррелированны как друг с другом, так и с оценками переменных n_a, n_e, $^{S}n_e$ и H_e. Для D'_{β}-оценок характерна более высокая вариативность (20,7 % против 8,9 % для D'_{α} и D'_{γ}).

Коэффициенты ранговой корреляции Кендалла (τ: тау) между переменными представлены в таблице 3. Из 91 τ-оценки статистически значимыми[5] по двухвыборочному тесту при доверительной вероятности 95 % было 46 (50,5 %). По множественному тесту с поправкой Бонферрони таковых было 19 (20,9 %), именно: между оценками α-разнообразия (кроме, H_o и n_a с n_e), индексами фиксации и четырьмя показателями аллельной дифференциации на базе гетерозиготности ($G''_{ST(NEI)}$, $F'_{ST(W\&C)}$, D_{JOST}, D_{CHAO}, кроме $G''_{ST(NEI)}$ с D_{JOST} и D_{CHAO}). Корреляции шенноновских оценок дифференциации (D'_{β}) со всеми другими показателями разнообразия были статистически незначимыми; без учёта поправки Бонферрони – значимая только с H_e и D'_{γ} (τ = 0,49). Шенноновская мера *γ*-разнообразия положительно, высоко и статистически значимо коррелировала с показателями *α*-разнообразия и ожидаемой гетерозиготностью (с поправкой Бонферрони), но отрицательно с индексами фиксации (без поправки).

[5]Априорный критический уровень статистической значимости (α) – вероятность ошибиться отказавшись от нулевой гипотезы (H_0). Может быть принят равным 0,001, 0,01, 0,05, 0,1. Обычный 5%-ный критический уровень (α = 0,05,) указывает, что ошибка в принятом решении возможна в 5 % случаев. Доверительная вероятность (P = 1-α) – надёжность, достоверность оценки. Рассчитанный по исходным данным *достигнутый* уровень статистической значимости есть «p_{value}». При справедливости H_0 pvalue> α = 0,05. При pvalue ≤ α = 0,05 H_0 отклоняется (разность средних ≠ 0, корреляция ≠0, есть влияние фактора и т.п.). Также используют: pvalue ≥ 0,1 – различие считается не значимым; 0,05 ≤ pvalue < 0,1 – слабозначимым; 0,01 ≤ pvalue < 0,05 – статистически значимым; 0,001 ≤ pvalue < 0,01 – сильнозначимым; pvalue < 0,001 – высокозначимым

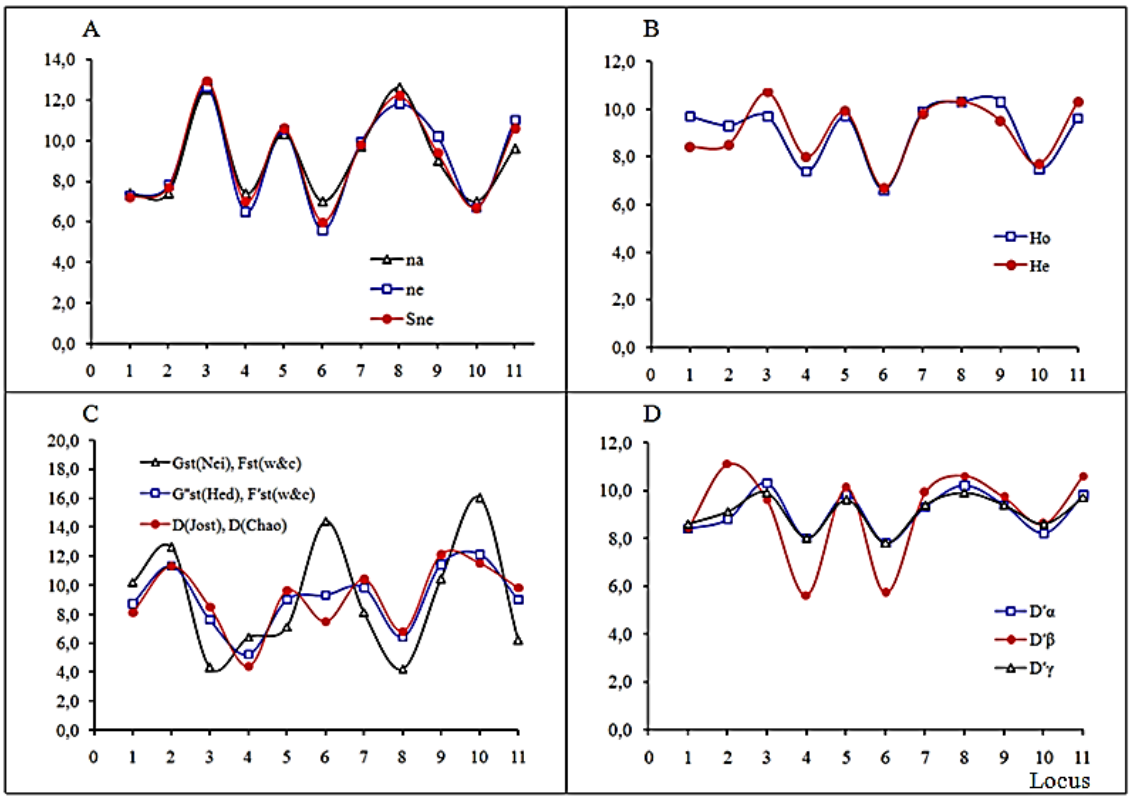

Рис. 1. **Профили полокусных оценок разнообразия/дифференциации** (по **Data2**): **A** – число аллелей на локус: n_a – фактических, n_e – эффективных по гетерозиготности, $^{S}n_e$ – эффективных по энтропии; **B** – гетерозиготность: H_o – наблюдаемая, H_e – ожидаемая; **C** – индексы фиксации ($G_{ST(NEI)}$ – по Нею, $F_{ST(W\&C)}$ – по Кокерхэм и Вейру) и меры дифференциации ($G''_{ST(NEI)}$, $F'_{ST(W\&C)}$, D_{JOST}, D_{CHAO}). **D** – меры аллельного разнообразия Шеннона/Шервина: D'_{α} – внутри выборок, D'_{β} – между выборками, D'_{γ} – общее /

Fig. 1. **Locus profiles based on diversity/differentiation estimates** (by **Data2**): Notes. **A** – the number of alleles per locus: n_a – actual, n_e – effective in heterozygosity, $^{S}n_e$ – effective in entropy; **B** – heterozygosity: H_o – actual, H_e – expected; **C** – fixation index ($G_{ST(NEI)}$ – by Hei′s, $F_{ST(W\&C)}$ – by Cockerham/Weir′s) and differentiation measures ($G''_{ST(HED)}$, $F'_{ST(W\&C)}$, D_{JOST}, D_{CHAO}. D – measures of Shannon/Sherwin′s allelic diversity: D'_{α} – within samples (breeds), D'_{β} – between samples (breeds), D'_{γ} – total

Дисперсионно-ковариационный анализ. Для выявления наличия статистически значимых различий между локусами по показателям аллельного разнообразия был проведён ANOVA (ANalysis Of VAriance) по **Data2**. Результаты тестов Shapiro-Wilk на нормальность распределения и Levene на равенство дисперсий свидетельствовали о *невозможности* применения параметрического ANOVA. Поэтому был использован непараметрический однофакторный ANOVA [28], в основе которого проверка H_0 о равенстве медиан сравниваемых групп (локусов); альтернативная гипотеза: по крайней мере, одна из медиан отлична от других.

Непараметрический ANOVA отклонил нулевую гипотезу (Test statistic = 40, p_{value}<<0,000). Медианный тест Муда (Mood's median test) подтвердил неравенство медиан (Test statistic = 38,8, p_{value} << 0,000). Оба критерия позволили выявить наличие различий между профилями показателей разнообразия локусов *в целом*, но не указали, между какими из них эти различия были, а между какими – нет.

Наиболее простой способ определения статистической значимости попарных сравнений – визуальный, именно: перекрываются или нет доверительные интервалы медиан на графиках «ящик с усами» (Box-and-Whisker Plot). Эти графики (рис. 2), построенные по **Data2**, визулизировали описательную статистику по каждому STR-локусу, именно: разброс оценок переменных («усы»), среднее (+) и медиана (горизонтальная линия), 25 и 75 процентили, определяющие квартильный размах – диапазон вокруг медианы («ящик»), который содержит 50 % оценок. Если интервальные вырезы по бокам «ящиков» двух локусов перекрываются, то различие медиан считается статистически незначимым (например, локусы 2 и 3; p_{value} > 0,05) и наоборот (локусы 2 и 4; p_{value} < 0,05). Хорошо видна статистическая гомогенность показателей разнообразия локусов 1, 5, 7, 11 и повышенная гетерогенность оценок локусов 8 и 10.

Таблица 3 – **τ-корреляции между показателями разнообразия** (по **Data1**)/
Table 3 – **τ-correlations between diversity indicators** (by **Data1**)

Variable	n_a	n_e	$^{s}n_e$	H_o	H_e	$G_{ST(NEI)}$	$F_{ST(WC)}$	$G''_{ST(H)}$	$F'_{ST(WC)}$	D_{JOST}	D_{CHAO}	D'_α	D'_β	D'_γ
n_a		0,0009	*0,0002*	0,0093	*0,0003*	0,0016	0,0028	0,1245	0,2533	0,4669	0,4669	*0,0005*	0,1245	*0,0005*
n_e	0,77		*0,0001*	0,0122	*0,0001*	0,0102	0,0064	0,3115	0,5297	0,9380	0,9380	*0,0000*	0,0516	*0,0000*
$^{s}n_e$	***0,88***	***0,92***		0,0223	*0,0000*	0,0047	0,0028	0,2087	0,3831	0,7533	0,7533	*0,0001*	0,0593	*0,0002*
H_o	0,61	0,585	0,534		0,0339	0,1693	0,2253	0,8084	0,8704	0,6861	0,6861	0,0076	0,2253	0,0076
H_e	***0,84***	***0,92***	***0,98***	0,49		0,0047	0,0028	0,2087	0,3831	0,8751	0,8751	*0,0002*	0,0411	*0,0002*
$G_{ST(NEI)}$	-0,74	-0,60	-0,66	-0,32	-0,66		*0,0001*	0,0064	0,0184	0,1021	0,1021	0,0158	0,3918	0,0064
$F_{ST(W\&C)}$	-0,70	-0,64	-0,70	-0,28	-0,70	***0,89***		0,0102	0,0278	0,1391	0,1391	0,0102	0,4835	0,0102
$G''_{ST(Hed)}$	-0,36	-0,24	-0,29	-0,06	-0,29	0,64	0,60		*0,0002*	0,0014	0,0014	0,3918	0,6971	0,2429
$F'_{ST(W\&C)}$	-0,27	-0,15	-0,20	0,04	-0,20	0,55	0,51	***0,88***		*0,0003*	*0,0003*	0,6374	0,3458	0,4321
D_{Jost}	-0,17	0,02	-0,07	0,09	-0,04	0,38	0,35	0,75	***0,84***		*0,0000*	0,9380	0,1391	0,9380
D_{Chao}	-0,17	0,02	-0,07	0,09	-0,04	0,38	0,35	0,75	***0,84***	***1,00***		0,9380	0,1391	0,9380
D'_α	***0,81***	***0,96***	***0,92***	0,62	***0,88***	-0,56	-0,60	-0,20	-0,11	-0,02	-0,02		0,0734	*0,0001*
D'_β	0,36	0,45	0,44	0,28	0,48	-0,20	-0,16	0,09	0,22	0,35	0,35	0,42		0,0356
D'_γ	***0,81***	***0,96***	***0,88***	0,62	***0,88***	-0,64	-0,60	-0,27	-0,18	-0,02	-0,02	***0,93***	0,49	

Примечания. Под диагональю – τ-оценки, над диагональю их p_{value}. Затонированы статистически **незначимые** τ-оценки при α = 0,05 по двухвыборочному тесту. Полужирным курсивом выделены статистически **значимые** τ-оценки по множественному тесту при α_{Bonf} = 0,05/91 = 0,0005. Здесь и далее p_{value} = 0,0000 есть p_{value} < 0,0001 /

Notes. Below the diagonal – τ-estimates, above the diagonal of their p_{value}. Gray tone –statistically not significant τ-estimates at α = 0.05 on a two-sample test. Statistically significant τ-estimates on the multiple test with α_{Bonf} = 0,05/91 = 0.0005 are highlighted in bold. Here and further p_{value} = 0.0000 is p_{value} < 0.0001/

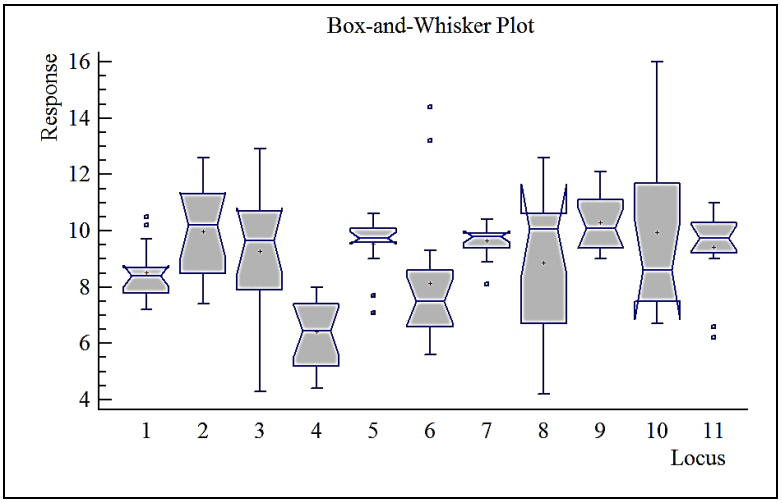

Рис. 2. **Графики описательной статистики для 11 STR-локусов** /
Fig. 2. **Graphs of descriptive Statistics for 11 STR loci**

Сравнение графиков «Box-and-Whisker» – это приближенный тест на статистическую значимость различия медиан. Более точную оценку дает непараметрический W-тест Mann-Whitney-Wilcoxon. В таблице 4 представлены уровни статистической значимости (p_{value}) при двухвыборочном (критерий Стьюдента) и множественном (предпочтительно) сравнении медиан локусов. В первом случае критический уровень статистической значимости был традиционный – α = 0,05. Во втором случае параметр α был модифицирован по Бонферрони: α_{Bonf} = 0,05/55 = 0,0009 (55 – число парных сравнений).

По Стьюденту H_0 отвергалась в 19 случаях (34,5 %), с поправкой Бонферрони – только в 8 случаях (14,5 %). Статистически значимыми были различия медиан локуса 1 с локусами 4 и 9, локуса 4 с локусами 1, 2, 5, 7, 9, 10 и 11.

Коэффициенты τ-корреляций между профилями локусов, рассчитанные по **Data2**, представлены в таблице 5.

Таблица 4 – **Уровни достигнутой статистической значимости (p_{value}) различий медиан STR-локусов (по Data2) /**
Table 4 – **Levels of achieved statistical significance (p_{value}) of differences in median STR-loci (by Data2)**

Locus	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1		0,0228	0,1540	0,0002	0,0256	0,0886	0,0057	0,7129	0,0006	0,5345	0,0272
2	0,0228		0,6457	0,0001	0,5196	0,0168	0,7823	0,2600	0,5346	0,7130	0,4481
3	0,1540	0,6457		0,0026	0,8901	0,0934	0,7122	0,7650	0,3458	0,9817	0,8902
4	0,0002	0,0001	0,0026		0,0000	0,0624	0,0000	0,0324	0,0000	0,0007	0,0001
5	0,0256	0,5196	0,8901	0,0000		0,0071	0,9815	0,9816	0,2898	0,8181	1,0000
6	0,0886	0,0168	0,0934	0,0624	0,0071		0,0024	0,4620	0,0016	0,0802	0,0215
7	0,0057	0,7823	0,7122	0,0000	0,9815	0,0024		0,7820	0,1737	0,6287	0,8897
8	0,7129	0,2600	0,7650	0,0324	0,9816	0,4620	0,7820		0,4342	0,4480	0,8902
9	0,0006	0,5346	0,3458	0,0000	0,2898	0,0016	0,1737	0,4342		0,3577	0,2695
10	0,5345	0,7130	0,9817	0,0007	0,8181	0,0802	0,6287	0,4480	0,3577		0,9450
11	0,0272	0,4481	0,8902	0,0001	1,0000	0,0215	0,8897	0,8902	0,2695	0,9450	

Примечания. Над диагональю – двухвыборочный тест Стьюдента (α = 0,05), под диагональю – множественный тест Бонферрони (α_{Bonf} = 0,05/55 = 0,0009). Затонированы статистически незначимые сравнения /

Notes. Above the diagonal – a two-sample Student′s test (α = 0,05), below the diagonal – a multiple Bonferroni test (α_{Bonf} = 0.05/55 = 0.0009). Gray tone – statistically not significant comparisons.

Таблица 5 – **Оценки τ-корреляций между профилями STR-локусов** (по **Data2**)/
Table 5 – **Estimates of τ-correlations between STR-loci profiles** (by **Data2**)

Locus	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1		0,0078	0,0014	0,9541	0,0024	0,0039	0,1569	0,0096	0,4254	0,0078	0,0014
2	0,53		*0,0000*	0,0305	*0,0002*	0,0085	0,8611	*0,0000*	0,0049	*0,0000*	0,0096
3	-0,64	***-0,90***		0,0911	*0,0002*	0,0016	0,7293	*0,0002*	0,0194	*0,0000*	0,0018
4	-0,01	-0,43	0,34		0,4201	0,8644	0,0207	0,1928	*0,0009*	0,2103	0,6469
5	-0,61	***-0,74***	***0,75***	0,16		0,0010	0,2884	*0,0000*	0,0876	*0,0001*	0,0031
6	0,58	0,53	-0,63	-0,03	-0,66		0,0267	*0,0006*	0,4643	*0,0004*	*0,0001*
7	-0,28	0,04	0,07	-0,46	0,21	-0,44		0,5235	0,0300	0,4143	0,0691
8	-0,52	***-0,83***	***0,75***	0,26	***0,84***	***-0,69***	0,13		0,0290	*0,0000*	0,0061
9	0,16	0,56	-0,47	***-0,67***	-0,34	0,15	0,44	-0,44		0,0556	0,5335
10	0,53	***0,82***	***-0,88***	-0,25	***-0,76***	***0,71***	-0,16	***-0,85***	0,38		0,0069
11	-0,64	-0,52	0,63	0,09	0,59	***-0,77***	0,36	0,55	-0,13	-0,54	

Примечания. Под диагональю – τ-оценки; над диагональю – p_{value} при критическом значении α = 0,05. Затонированы статистически незначимые значения. Курсивом выделены статистически значимые корреляции при α_{Bonf} = 0,05/55 = 0,0009 /

Notes. Under the diagonal – τ-estimates; above the diagonal – p_{value} at critical value α = 0.05. Gray tone – statistically not significant τ-values. Italics indicate a statistically significant correlations at α_{Bonf} = 0.05/55 = 0.0009

По множественному тесту Бонферрони статистически значимыми было 14 (25,5 %) τ-оценок. Положительные корреляции имели место между локусами 2 и 10, 3 и 5, 3 и 8, 5 и 8, 6 и 10 (τ = 0,71-0,84); отрицательные – локуса 2 с локусами 3, 5 и 8; между локусами 3 и 10, 4 и 9, 5 и 10, локуса 6 с локусами 8 и 11, между локусами 8 и 10 (τ = -0,9…-0,67; при τ = -1 ранги коррелируемых пар располагаются во взаимно обратном порядке).

Анализ главных компонент. Представленные в таблицах 1 и 2 оценки по локусам можно рассматривать как матрицы многомерных наборов данных[6] типа «объект-переменные» размерностью 11×14, где объектами являлись 11 локусов, а переменными/признаками – 14 оценок, характеризующие разные аспекты аллельного разнообразия. Следовательно, эти данные можно представить не как массивы чисел (и отдельные одномерные графики (рис. 1)),

[6]Многомерные данные – это набор данных, в котором для каждого объекта (наблюдения) содержится информация о трёх или более его характеристиках (показателях, признаках). Для анализа таких данных используют методы многомерной статистики.

а как векторы в многомерном пространстве. Тогда, набор данных – это облако точек в пространстве большой размерности. Объекты-локусы в 14-мерном пространстве реально невозможно графически отобразить и выявить наличие каких-либо отклонений от случайного рассеивания (закономерности). Для снижения размерности использовали методы многомерной статистики: анализ главных компонент (Principal Components Analysis, **PCA**), неметрическое многомерное шкалирование (Non-metric MultiDimensional Scaling, **nMDS**), кластерный анализ (Cluster Analysis, **CA**).

PCA часто считают формой факторного анализа. Суть PCA в том, чтобы выявить существование скрытых (латентных) факторов или компонент. При этом компоненты строятся в порядке убывания объясняемой ими доли суммарной дисперсии исходных переменных. Это позволяет ограничиться первыми (главными) компонентами («новыми» или «синтетическими» латентными переменными). Главные компоненты определяются как линейные комбинации исходных переменных, некоррелированные друг с другом и захватывающие максимальную долю дисперсии данных. Если исходные переменные коррелируют, то 2-3 главных компонента может быть достаточно для описания бóльшей части общей дисперсии. Процедура PCA исследует связь между переменными, извлекает главные компоненты из имеющегося числа переменных (снижает размерность данных) и проецирует объекты в пространство главных компонент, т. е. определяет ординацию объектов[7]. Кроме того, компонентные оценки (scores), как новые комплексные (интегрированные) переменные, отражающие реальную структуру объектов и наиболее полно передающие исходную информацию (относительно, например, агрегирования путём суммирования), могут быть использованы в иных видах анализов (множественная регрессия, кластерный анализ и др.).

Процедура PCA на начальном этапе создаёт график «каменистой осыпи» (screeplot), с помощью которого определяют число главных компонент. Такой график по **Data1** представлен на рисунке 3, А. Собственное значение (eigenvalue) – это вклад компоненты в общую изменчивость данных. Вторая точка на рисунке 3, А, где начинается резкий спад кривой, соответствует числу главных компонент (PC1 и PC2) при условии сохранения максимальной дисперсии. Незначимые компоненты (PC3-PC14) находятся далее на максимально замедленной части кривой (т.н. «щебень»). Собственное значение, деленное на число переменных, отражает долю дисперсии, которая соответствует данной компоненте. Эту долю дисперсии интерпретируют как показатель информативности (мощности) компоненты. Сумма дисперсий по главным компонентам – показатель того, насколько полно выделяемые компоненты представляют анализируемый набор данных (подобие коэффициента детерминации).

Главные компоненты представляют собой линейные комбинации переменных, где «весами» являются «нагрузки» (loading). Компонентная нагрузка (аналог коэффициента корреляции) отражает связь между переменной и компонентой; имеет диапазон [-1; +1]. Чем больше величина нагрузки, тем сильнее связь (влияние) переменной с компонентой (рис. 3, В). Квадрат компонентной нагрузки (аналог частного коэффициента детерминации) интерпретируют как часть дисперсии переменной, объясняемая данной компонентой. Сумма квадратов всех компонентных нагрузок по переменной равна 1 – полной дисперсии переменной. Сумма квадратов всех нагрузок по компоненте (= собственному значению) делённая на число переменных равна доле дисперсии соответствующей компоненты.

Каждой компоненте может быть присвоено имя по переменным с наибольшими нагрузками. На рисунке 3, В видно, что переменные n_a, n_e, ^Sn_e, H_o, H_e, D'_α, D'_γ, характеризующие внутривыборочное аллельное разнообразие, имели бóльшие нагрузки по первой компоненте (PC1). Поэтому PC1 была определена как «альфа-компонента». По PC2 бóльшие нагрузки имели переменные $G''_{ST(HED)}$, $F'_{ST(W\&C)}$, D_{JOST}, D_{CHAO} и D'_β, которые характеризовали разнообразие между выборками (породами). PC2 была названа «бета-компонентой». Выделенные две главные компоненты (PC1 и PC2) или *латентные переменные* (αLV и βLV) были однозначно определены, соотносясь с *α*- и *β*-разнообразием, составляющих общее аллельное разнообразие STR-локусов объединённых породных выборок.

[7]Ординация – собирательное понятие для многомерных методов обработки данных. Ординация в узком смысле представляет собой нахождение таких координатных осей на плоскости, относительно которых можно выполнить оптимальное проецирование многомерных анализируемых объектов. Ординация позволяет расположить объекты вдоль некоторых осей, основываясь на значениях переменных, исследовать вариабельность объектов, их близость/ отдалённость и наличие структуры, т. е. выделить группы объектов и визуализировать их в 2-3-мерном пространстве. Ординация и классификация (кластеризация) – два метода, которые, в некоторой степени, дополняют друг друга.

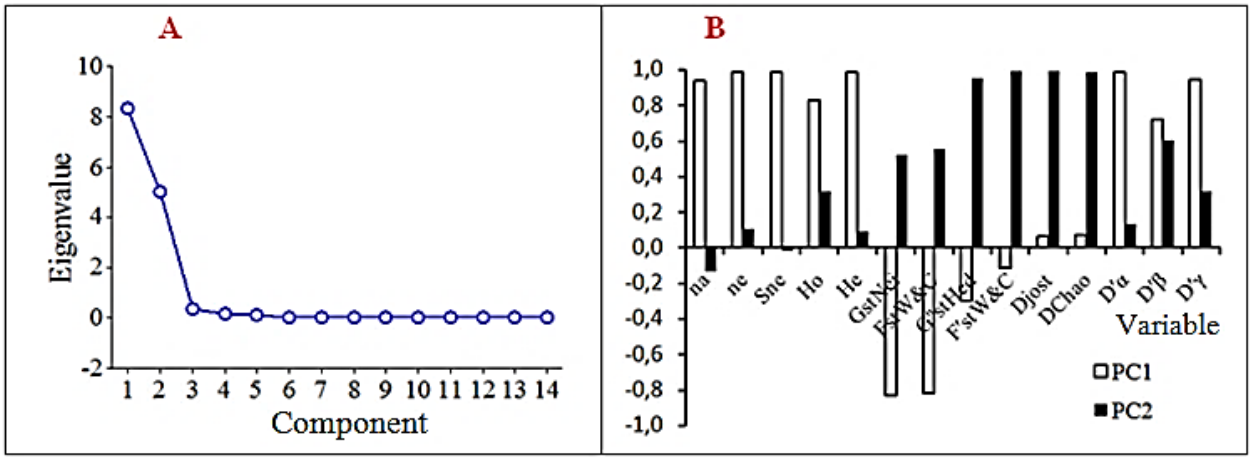

Рис. 3. **Графики собственных значений компонент (A) и их нагрузок (B) /**
Fig. 3. **Graphs of eigenvalues of components (A) and their loads (B)**

С помощью программы PCA рассчитали компонентные оценки (координаты локусов в двумерном пространстве) и отобразили в ортогональной системе координат структурные особенности изучаемых 11 STR-локусов в форме графической проекции – ординации (рис. 4, слева).

PC1 объясняла 59,4 % общей дисперсии исходных данных, PC2 – 35,8 %. Суммарная информативность (мощность) PCA составила 95,2 %[8]. Компоненты PC3-PC14, которые были игнорированы, имели вклад в общую дисперсию 4,8 %. Преобразование исходной матрицы данных размерностью 11×14 в матрицу компонентных координат локусов размерностью 11×2 произошло без существенной потери информации. Две латентные переменные характеризовали общую дисперсию почти также хорошо, как исходные 14, но при этом были ортогональны (независимы) друг от друга. Отметим, величина дисперсии, которая объяснялась PC2 (βLV), равная 35,8 %, была близка к сводным оценкам показателей дифференциации породных выборок (табл. 1), особенно к $G''_{ST(HED)}$ = 37,5%, $F'_{ST(W\&C)}$ = 35,1 % и к шенноновской относительной оценке *β*-разнообразия – 36,2 %.

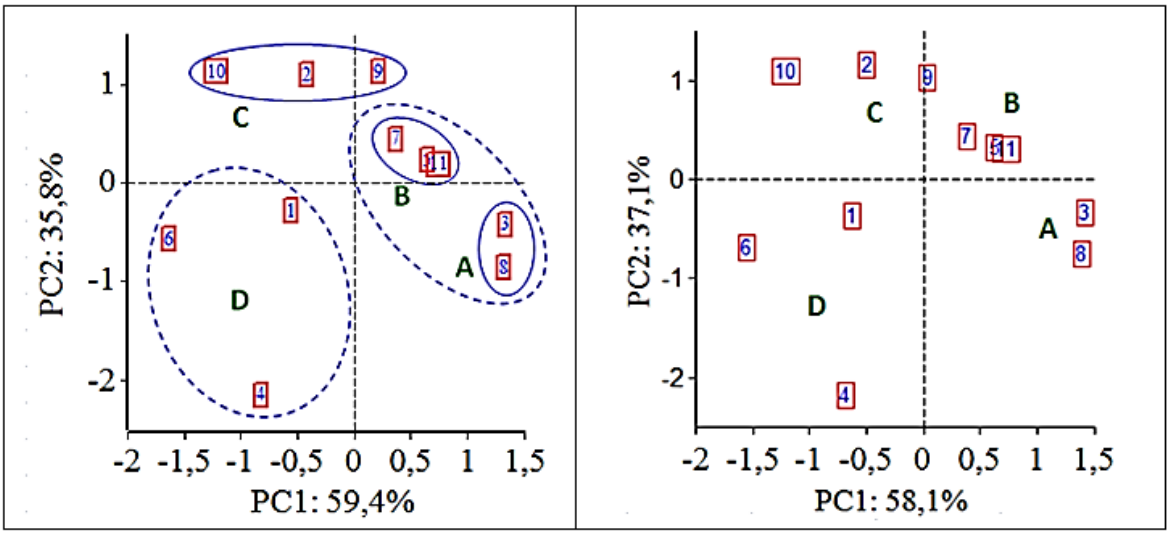

Рис. 4. **Ординация STR-локусов в пространстве главных компонент (Слева – PCA по Data1 [11×14], справа – PCA по Data1r [11×7]) /**
Fig. 4. **Ordination of STR loci in the space of principal components (left – PCA by Data1 [11×14], right – PCA by Data1r [11×7])**

[8]PCA считают успешным, если суммарная объяснённая дисперсия (информативность) главных компонент не ниже 70%. Чем меньше информативность компонент, тем менее надёжны компонентные оценки (координаты объектов), тем ниже подтверждение правильности (валидности) ординации.

Аппроксимация с использованием только двух латентных переменных позволила обобщить результаты в виде точечного графика – ординации. На ординации (рис. 4, слева) выделились группы локусов, близкие по координатным оценкам: локусы 3 и 8 – группа A, локусы 5, 11 и 7 – группа B, локусы 10, 2 и 9 – группа C. Локусы условной группы D (4, 6, 1) были удалены как друг от друга, так и от других групп. Поэтому определены как «нетипичные». Локус 6 и особенно 4 подходили под определение «выбросы».

Локусы группы A имели самые высокие значения αLV и ниже среднего βLV ($\alpha^{++}\beta^{-}$). Локусы группы B относились к категории «$\alpha^{+}\beta^{+}$», а локусы группы C – к «$\alpha^{-}\beta^{++}$». В группе D локус 1 имел величины LV ниже среднего ($\alpha^{-}\beta^{-}$), а локусы 4 и 6 – предельно низкие ($\alpha^{--}\beta^{--}$).

PCA с редуцированными данными. Качество, устойчивость ординации (валидность[9]) проверяется, как правило, согласованностью её с таковыми, полученными на иных наборах данных и/или с использованием других методов многомерной статистики.

Набор исходных данных размерностью 11×14 (**Data1**) сократили до размерности 11×7 путём удаления «избыточных» переменных одного типа (**Data1r**). Были оставлены переменные n_a, $^{S}n_e$, H_e, $G_{ST(NEI)}$, $F'_{ST(W\&C)}$, D_{CHAO} и D'_{β} (в табл. 1 и 2 помечены «*»). Три первые характеризовали α-разнообразие, остальные – β-разнообразие. PCA-ординация по **Data1r** представлена на рисунке 4, справа.

Сокращение числа исходных переменных в два раза не отразилось на общей информативности PCAr (95,2 %). Незначительно снизилась дисперсия PC1, но настолько же повысилась дисперсия PC2. Не отразилось сокращение числа переменных и на ординации локусов – также можно было выделить группы A, B, C и D. Для оценки согласованности (сходства) ординаций использовали анализ Прокруста[10] – PROTEST[11] [26], который выдает m_{12}-статистику, называемую «дистанцией Прокруста» (чем меньше величина, тем лучше), и вероятность отвержения нулевой гипотезы (статистически значимое соответствие, если $p_{perm} \leq 0{,}05$). Приближённую величину сходства (r^2) можно получить из отношения $m_{12} \approx 1 - r^2$ [25]: $r^2 \approx 1 - m_{12}$. PROTEST на основе пермутации компонентных оценок (999 перестановок) показал статистически высокозначимое сходство PCA- и PCAr-ординаций, именно: $m_{12} = 0{,}0034$, $p_{perm} = 0{,}0001$; $r^2 = 0{,}997$.

Неметрическое многомерное шкалирование (nMDS). В контексте проверки устойчивости PCA-ординации был проведён nMDS-анализ, причём по преобразованным полокусным оценкам отобранных переменных (**Data2r**; табл. 2 с «*»). Суть nMDS состоит в нелинейном преобразовании дистанций между объектами. Процедура nMDS пытается как можно точнее представить парные различия между объектами в низкоразмерном пространстве. С помощью алгоритма nMDS рассчитывали дистанции между объектами, присваивали им ранги и размещали в пространстве 2-3-мерной системы координат (шкал) таким образом, чтобы ранги дистанций в результирующей ординации воспроизводились с сохранением порядка исходных значений. Например, если исходная дистанция между объектами 5 и 8 имела ранг 16 (из всех дистанций между любыми двумя объектами), то на графике ординации расстояние между точками 5 и 8 в идеале должно оставаться по-прежнему 16-м по величине.

Критерием качества nMDS-ординации является стресс (stress) – мера отклонения финальной конфигурации объектов от исходных дистанций (в контексте требования рангового соответствия). Алгоритм nMDS направлен на нахождение оценок координат объектов, минимизирующих величину стресса. Ординация при стрессе ≥0,3 считается случайной, 0,2 – плохой, 0,1 – удовлетворительной, 0,05 – хорошей, 0,025 – отличной и 0,000 – идеальной. Дополнительно вычисляется коэффициент R^2 (детерминации), который показывает долю дисперсии исходных различий, учтённую выделенными шкалами (осями; в PCA – компоненты). Чем ближе R^2 к 1, тем полнее данные шкалы воспроизводят исходные различия между объектами.

[9]Валидность – обоснованность и пригодность применения методик и результатов исследования в конкретных условиях. Валидация – процедура, дающая высокую степень уверенности в том, что конкретный процесс, метод или система обладает повторяемостью (воспроизводимостью и устойчивостью).

[10]Прокрустов анализ (Procrustes Analysis) – статистический метод, который сравнивает наборы многомерных форм, пытаясь преобразовать их в состояние суперналожения. В программе PROTEST это достигается путем минимизации сумм квадратов расстояний между соответствующими точками в каждой форме посредством перемещения, отражения, вращения и масштабирования их координатных матриц.

[11]STATGRAPHICS® Centurion XVI User Manual. 2010.

На рисунке 5 слева представлена nMDS-ординация STR-локусов по матрице Gower-дистанций. Под шкалами даны величины R^2. Их суммарное значение составило 70,4 + 20,6 = 91 %. Это на 4,2 процентных пункта ниже, чем информативность в двух предыдущих анализах. Значение суммарной R^2 и стресса, равного 0,032, характеризовали качество nMDS-ординации как «достаточно хорошее». График Шепарда (Shepard plot; рис. 5 справа) иллюстрирует различия между полокусными исходными и порядковыми дистанциями. Расхождения были незначительными, что свидетельствовало о хорошей линейной зависимости (корреляции) и заслуживающем доверия визуализации отношений между локусами. PROTEST-анализ показал высокое соответствие nMDS-ординации с PCA-ординацией по **Data1** (m_{12} = 0,044, p_{perm} = 0,001; r^2 = 0,956).

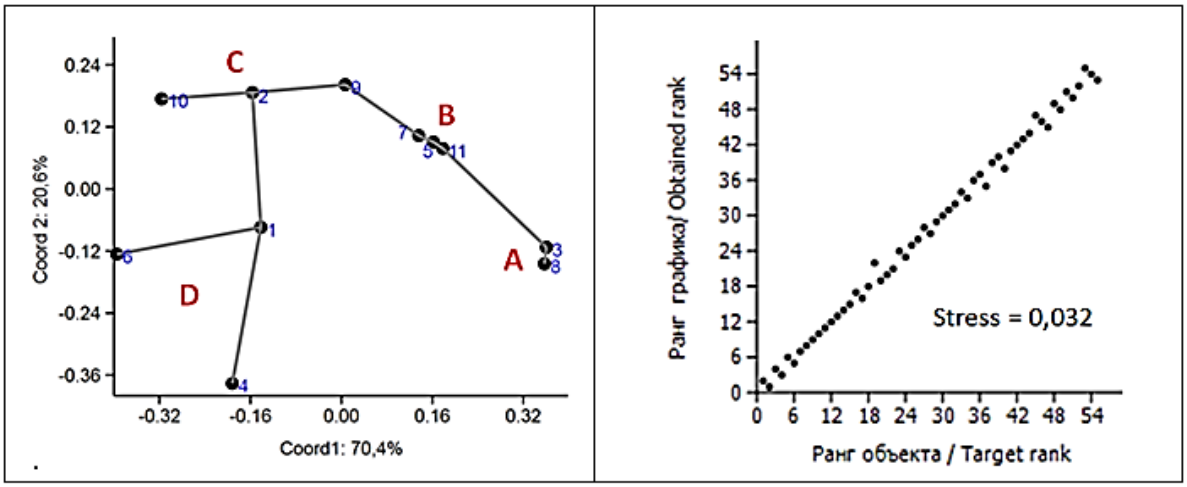

Рис. 5. **nMDS-ординация STR-локусов (слева) и график Шепарда (справа) качества подгонки модели (по Data2r)** /
Fig. 5. **nMDS-ordination STR-loci (left) and Shepard plot (right) of model fit quality (by Data2r)**

Вероятные цепочки отношений между локусами визуализировал некорневой граф (рис. 5 слева – линии между точками-локусами), построенный методом «минимального остовного древа» (Minimum Spanning Tree, MST). MST использовали в качестве вспомогательного средства, способствующего корректной группировке локусов. Последняя (по преобразованным данным) была аналогичной таковым в обоих PCA. Во всех трёх ординациях выделялись «выбросы» – локусы 4 и 6, удалённые от групп A, B и C на значительном расстоянии.

Межлокусные Gower-дистанции. Ординация по nMDS (как и по PCA) – это «пространственная карта» локусов, но без указания расстояний. Близость/удалённость локусов была измерена Gower-дистанцией[12] (d_G), исходя из различий профилей локусов – векторов полокусных оценок (табл. 6).

Значения d_G-оценок находились в диапазоне 0,06-0,74 со средней 0,38. Тест Shapiro-Wilk показал большую вероятность того, что d_G-оценки подчинялись закону нормального распределения (W = 0,966, p_{value} = 0,118). При классификации d_G (с величиной класса 0,2) в первую группу с интервалом 0,064÷0,194 и средним **0,13** вошло 8 дистанций (14,5 %), во вторую группу (0,223-0,396, **0,29**) – 22 (40 %), в третью (0,402-0,576, **0,51**) – 20 (36,4 %), четвёртую (0,624-0,736, **0,69**) – 5 дистанций (9,1 %). Почти половина d_G-оценок была в диапазоне 0,4-0,7. Похожие результаты наблюдали при использовании метода максимального правдоподобия (модуль Mixture analysis программы PAST3): **{G1:** n = 5 (9,1 %), mean = **0,09**}, {**G2**: 27 (49,1 %), **0,28**}, {**G3:** 19 (34,5%), **0,53**} и **{G4**: 4 (7,3 %), **0,70**}. Между локусами группы B усреднённая дистанция составила 0,08, группы A – 0,12, группы C – 0,22. Между локусом 6 и локусами 2, 5 и 8 дистанции были 0,4, 0,56 и 0,7; локуса 4 с предыдущими – 0,51, 0,55 и 0,5 соответственно. Самая большая дистанция имела место между локусами 8 и 10 – 0,74.

[12]Gower-дистанция измеряет среднее различие по всем переменным, которые нормализованы по рангам каждой переменной: $d_{jk} = (1/n) \Sigma_m |x_{jm} - x_{km}| / ({}^{max}x_{sm} - {}^{min}x_{sm})$, где d_{jk} – дистанция между j и k локусами; n – число переменных; x_{jm} и x_{km} – m-я переменная в j и k локусах; ${}^{max}x_{sm}$ и ${}^{min}x_{sm}$ – максимальное и минимальное значения m-ой переменной в выборке.

Таблица 6 – **Gower-дистанции между профилями STR-локусов (по Data2r) /**
Table 6 – **Gower-distances between STR-locus profiles (by Data2r)**

Locus	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1		-	-	-	-	-	-	-	-	-	-
2	0,229		-	-	-	-	-	-	-	-	-
3	0,469	0,570		-	-	-	-	-	-	-	-
4	0,281	0,510	0,624		-	-	-	-	-	-	-
5	0,340	0,333	0,237	0,545		-	-	-	-	-	-
6	0,225	0,396	0,689	0,313	0,564		-	-	-	-	-
7	0,298	0,279	0,291	0,528	0,064	0,523		-	-	-	-
8	0,489	0,576	0,116	0,568	0,269	0,698	0,333		-	-	-
9	0,293	0,190	0,410	0,574	0,194	0,513	0,130	0,463		-	-
10	0,257	0,186	0,678	0,501	0,466	0,289	0,402	0,736	0,284		-
11	0,351	0,336	0,234	0,540	0,077	0,575	0,099	0,240	0,223	0,496	

Разброс локусов по α- и β-оценкам Шеннона/Шервина. Информационно-энтропийным методом Шеннона/Шервина рассчитываются *α*- и *β*-показатели, характеризующие аллельное разнообразие внутри и между популяциями. Их выборочные относительные оценки по локусам представлены в таблице 1 (D'_α и D'_β). Было получено рассеяние локусов в двумерной плоскости этих оценок (рис. 6), что соответствовало редукции **Data1** до размерности 11×2.

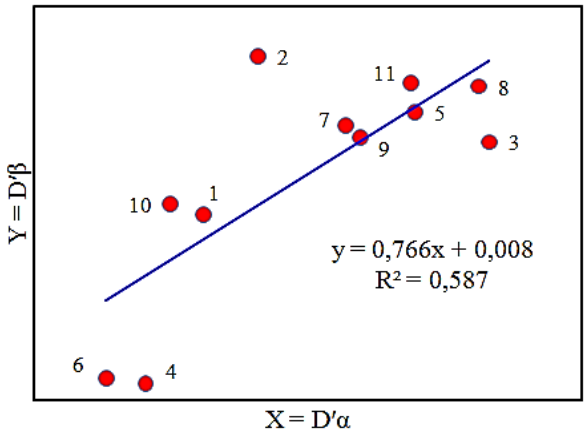

Рис. 6. **Разброс локусов на плоскости *α*- и *β*-оценок Шеннона/Шервина** /
Fig. 6. **The spread of loci by the *α*- and *β*-estimates of Shannon/Sherwin's**

Разброс локусов заметно отличался от таковых в PCA и nMDS ординациях (рис. 4 и 5). PROTEST выявил повышенные величины m_{12}-статистики с ординациями по PCA (0,284) и nMDS (0,221). Это определило r^2-оценки 0,72 и 0,78 соответственно. Тем не менее согласованность ординаций по шкале Чеддока подходила под категорию «сильная» и была статистически значимая ($p_{perm} \leq 0{,}01$).

Имела место парная группировка локусов: 3-8, 5-11, 7-9, 1-10 и 4-6. Две первые пары локусов характеризовались высокими показателями *α*- и *β*-разнообразия; в целом, соответствовали группам локусов A и B на ординациях рисунков 4 и 5. Три первые пары, в принципе, можно объединить в одну группу. Локусы двух последних пар характеризовались низкими показателями *α* и *β*-разнообразия и демонстрировали существенно бóльшую близость, чем на PCA- и nMDS-ординациях. Локус 2 находился в стороне от остальных и имел ниже среднего D'_α-оценку и самую высокую D'_β-оценку.

Регрессионный анализ показал статистически значимую связь между оценками переменных D'_α и D'_β (p_{value} = 0,006). Имела место тенденция повышения полокусных оценок межпородного разнообразия с увеличением оценок внутрипородного разнообразия (в PCA *α*- и *β*-латентные переменные были ортогональными). Коэффициент детерминации R^2 (рис. 6) свидетельствовал о том, что линейная модель объясняла 58,7 % вариабельности полокусных оценок межпородного разнообразия (D'_β). Коэффициент корреляции, равный 0,766 ($\sqrt{R^2}$), указывал на умеренно сильную взаимосвязь между D'_α и D'_β переменными.

Оценки внутрипородного разнообразия (D'_α) являлись производными от числа эффективных аллелей [по энтропии]. Последние были продуктами числа аллелей на локус (n_a) и их распространённости (частоты и выравненности частот). Проводили регрессионный анализ D'_β

на n_a; статистика R^2 составила 0,296, тогда 58,7 - 29,6 = 29,1 % изменчивости оценок D'_β объяснялось распространённостью аллелей. Следовательно, можно полагать, что число аллелей на локус и их распространённость в равной степени оказывали влияние на вариабельность оценок дифференциации породных выборок.

Кластерный анализ (CA). Стандартизированные компонентные оценки по локусам из PCA по **Data1r** были использованы в CA, как новые синтетические некоррелируемые переменные. CA – это процедура упорядочивания объектов (локусов) в сравнительно однородные группы (кластеры) на основе парного сравнения по предварительно определённым и измеренным критериям[13]. Использован алгоритм невзвешенного парно-группового метода с арифметическим усреднением (Unweighted Pair-Group Method using arithmetic Averages – UPGMA), евклидовой метрикой и построением древообразной структуры – дендрограммы, которая визуализировала отношения между локусами.

В UPGMA дистанция между попарно сравниваемыми объектами зависит от наблюдаемых различий по переменным. Пары объектов, между которыми дистанции минимальны, группируются (создают кластер) в первую очередь и оказываются на соседних ветвях дендрограммы. Межкластерная дистанция определяется как среднее всех парных дистанций между членами двух кластеров. Точки ветвления помещаются посередине между двумя объектами или кластерами. Расстояние между двумя объектами является суммой длин ветвей.

Дендрограмма (рис. 7) иллюстрирует последовательность выделения четырёх кластеров: A, B, C и D, состав которых был идентичен группам, полученным в PCA и nMDS (рис. 4, 5). Кофенетическая корреляция, как мера обоснованности дендрограммы [29], была достаточно высокой 0,74[14].

Для оценки адекватности построенного и реального древ используют численный ресэмплинг (метод бутстрэп). Бутстрэп имитирует формирование новых выборок путём повторного выбора (с возвращением) из исходного набора данных. К бутстрэп-выборкам применяли тот же алгоритм, что и к исходному набору данных. На основании множества бутстрэп-древ определяли выборочные свойства кластеринга. Бутстрэпинг 999 псевдовыборок показал 100%-ю вероятность объединения в кластер B локусов 5, 7 и 11. Бутстрэп-вероятность объединения локусов 3 и 8 (кластер A) была 73 %. Вероятность кластера C составила 73 %, а кластера D – 47 %.

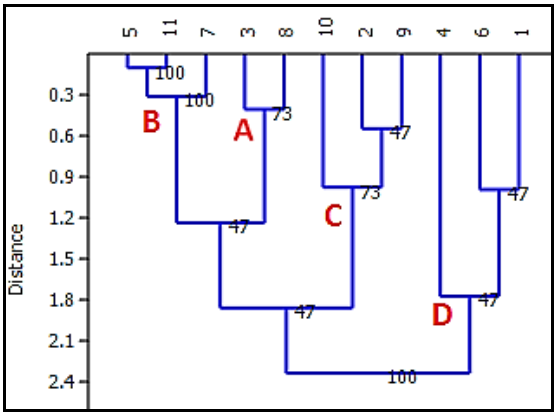

Рис. 7. **UPGMA-дендрограмма кластеринга локусов по PCA-оценкам** /
Fig. 7. **UPGMA-dendrogram of clustering of loci by PCA-scores**

Существует эмпирическое правило валидности дендрограммы, именно: устойчивая (достоверная) топология должна сохраняться при изменении метода кластеринга. Например, если результаты иерархического CA на 70 % и более совпадают с группировкой по методу k-средних, то предположение об устойчивости принимается. Расчёты по методу k-средних показали 100%-ное совпадение полученных кластеров.

Сходство локусов с центроидом. В методах ординации «центроид» (центральная точка) – это среднее (взвешенное) положение чего-либо в пространстве ординации, то есть вдоль всех осей ординации одновременно. Для определения центроида использовали вектор сводных оценок показателей разнообразия по всем локусам. Были рассчитаны евклидовы дистанции между локусами и центроидом (d_E) с их последующей стандартизацией (d'_E)[15]. Разброс локусов, стандарти-зированные дистанции и гистограмма сходства ($S = 1 - d'_E$) с центроидом представлены на рисунке 8.

[13]Результатом может быть формирование нескольких кластеров, в каждом из которых содержатся объекты, обнаружившие ранее неизвестные статистически значимые закономерности, взаимосвязи. Последующий анализ кластеров может выявить некоторые объективные характеристики, по которым эти кластеры различаются.
[14]Кофенетическая корреляция – оценка соответствия расстояний на дендрограмме расстояниям в исходном многомерном пространстве; насколько хорошо характер отношений (сходство/несходство) между объектами (локусами) представляется дендрограммой.
[15]Стандартизация по минимаксной формуле: $d'_E = (d_E - d_{E.min}) / (d_{E.max} - d_{E.min})$, где d_E – евклидова дистанция между локусами; $d_{E.min}$ и $d_{E.max}$ – минимальное и максимальное значения из всех дистанций.

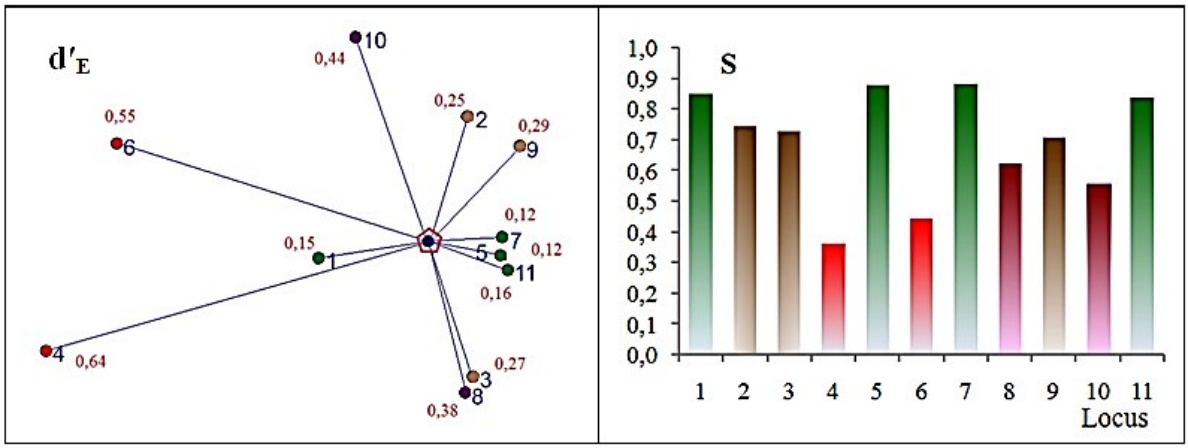

Рис. 8. **Дистанции** (d'_E) **и сходства** ($S = 1 - d'_E$) **STR-локусов с центроидом /**
Fig. 8. **Distances** (d'_E) **and similarities** ($S = 1 - d'_E$) **of STR-loci with centroid**

Локусы 4 и 6 имели показатели сходства с центроидом на уровне 40 %, локусы 8 и 10 – 60 %, локусы 2, 3 и 9 – 70-75 %. Наибольшее сходство с центроидом имели локусы 1, 5, 7 и 11 – S = 84-88 %. Можно полагать, что использование только этих четырёх локусов обеспечит получение вектора сводных оценок разнообразия/дифференциации, близкого к вектору сводных оценок по 11 локусам. Была сделана проверка этого предположения. Дополнительно рассчитаны показатели разнообразия/дифференциации по субвыборке, включающей только 4, 6, 8 и 10 локусы (S≤62 %). Ниже представлены результаты:

Locus	n_a	$^{S}n_e$	H_e, %	$G_{ST(NEI)}$, %	$F'_{ST(W\&C)}$, %	D_{CHAO}, %	D'_{β}, %
1÷11	5,0	3,7	66	10,3	35,1	30,8	36,2
1,5,7,11 (S≥84 %)	5,1	3,9	69	10,4	34,8	29,9	38,7
4,6,8,10 (S≤62 %)	4,7	3,3	59	11,6	31,7	24,3	30,5

Очевидно, что оценки по локусам с S ≥ 84 % ближе к «истинным» (по 11-ти локусам), чем по локусам с S ≤ 62 %. В первом случае расхождения были в пределах статистической ошибки – 3,4 % (mean absolute percentage error, MAPE[16]). Во втором случае MAPE составила 12,4 %, что можно интерпретировать как статистически значимое отклонение от «истинных» оценок.

γLV-оценки. Альфа- и бета-компонентные оценки локусов (координаты) являлись, как отмечалось, линейными комбинациями исходных мер разнообразия или новыми латентными переменными. Эти переменные, обозначаемые αLV и βLV, были независимыми, отражали реальную структуру взаимосвязей оригинальных переменных и наиболее полно передавали исходную информацию. Их можно сравнить с субиндексами в теории построения селекционных индексов Хендерсона [30], именно: αLV – «субиндекс» интегрированных переменных *α*-разнообразия, учитывающий переменные *β*-разнообразия, и βLV – «субиндекс» интегрированных переменных *β*-разнообразия, учитывающий переменные α-разнообразия. Так как переменные αLV и βLV ортогональны, то обобщённый «индекс» для k-го локуса рассчитывали по уравнению: $\gamma LV_k = (\alpha LV_k + \beta LV_k)/2$ («веса» для «субиндексов» равны 0,5). Получаемая в итоге новая агрегатная переменная (γLV_k) характеризовала k-й локус как *оценщика* тотального показателя разнообразия анализируемой выборки.

По **Data1r** для каждого локуса были рассчитаны γLV-оценки и сопоставлены с полокусными оценками *γ*-разнообразия Шеннона/ Шервина (D'_γ)[17]. Использовали непараметрический тест Колмогорова-Смирнова, который проверял любые различия распределений, любых параметров, без конкретизации каких именно. Критерий Колмогорова-Смирнова составил D = 0,36. Так как $p_{value(asymptotic)} = 0,47 > \alpha = 0,05$, то H_0 не отвергалась – между распределениями D'_γ и γLV не было выявлено никаких различий, разные методы «обработки» привели к практически однородным выборкам оценок.

[16]MAPE = $(100 \times \Sigma|e_i|/\theta^*)/n$, где $e_i = \theta_i - \theta^*$; θ_i – i-ая оценка, θ^* – «истинная» оценка.
[17]D'_γ-оценки были стандартизированы.

Синхронность вариативности полокусных D'$_\gamma$ и γLV оценок показана на рисунке 9 слева. Точечная оценка корреляции Пирсона, r = 0,92 (p_{value} = 0,00004), с вероятностью 95 % находилась в интервале 0,73-0,98. В то же время локусы 4 и 6 подходили под определение «выбросы», что могло повлиять на оценки и результаты проверки H_0. Ранговая τ-корреляция составила 0,67 (Upper Probability = 0,002 и Lower Probability = 0,998, p_{value} = 0,005). Если согласованность совместной изменчивости D'$_\gamma$ и γLV оценок была достаточно высокой (R^2 = 85 %; рис. 9 справа), то согласованность рангов подходила под категорию «умеренно-заметная». Ранги локусов по двум переменным были:

Locus	3	9	11	5	7	8	2	10	1	6	4
D'$_\gamma$-rank	2	5	3	4	6	1	7	9	8	11	10
γLV-rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**

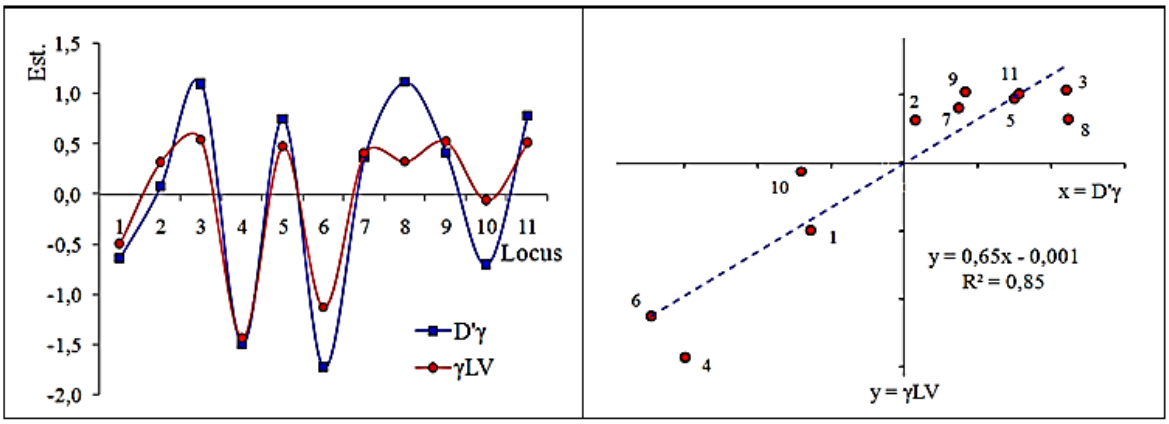

Рис. 9. **Рассеяние локусов по D'γ и γLV в одномерной (слева) и двумерной (справа) плоскостях /**
Fig. 9. **Scattering of loci by D'y and yLV in one-dimensional (left) and two-dimensional (right) planes**

Подобные ранги имели три локуса (27,3 %), повысили по γLV-переменной – пять (45,4 %), понизили – три (27,3 %). Шесть локусов расходились в рангах на один разряд, локус 9 повысил ранг с 5 до 2, а локус 8 снизил ранг с 1 до 6.

Локус 3, с рангом 1 по γLV, имел максимальные показатели по эффективному числу аллелей (n_e, ^Sn_e), ожидаемой гетерозиготности (H_e,), внутрипородному разнообразию Шеннона/Шервина (D′$_\alpha$), высокие, но не максимальные, показатели межпородной дифференциации и общего разнообразия (D′$_\beta$, D′$_\gamma$), минимальную величину индекса фиксации (табл. 1). В PCA-ординации локус 3 относился к группе А (рис. 4) и имел соответствие с финальной мультилокусной оценкой 73 % (рис. 8).

Локус 8, лучший по D'$_\gamma$ (по γLV ранг 6), имел максимальные значения по аллельному богатству (n_a; высокие по n_e, ^Sn_e), фактической гетерозиготности (H_o; высокое по H_e), γ-разнообразию Шеннона/Шервина (D′$_\gamma$; высокое по D′$_\alpha$) и минимальные значения индексов фиксации ($G_{ST(NEI)}$, $F_{ST(W\&C)}$). Как и локус 3 относился к группе А, но его сходство с центроидом составило 63 %.

Локус 4, худший по γLV и D'$_\gamma$, имел минимальные показатели по всем мерам межпородной дифференциации. По остальным показателям разнообразия величины были близки к минимальным. В PCA-ординации входил в группу «нетипичные» и имел сходство профиля с финальными оценками 36 %.

Заключение. По 11 STR-локусам ДНК 84 генотипированных быков семи пород были рассчитаны 14 показателей, характеризующих разные аспекты аллельного разнообразия объединённой выборки. Оригинальные и стандартизированные оценки сформировали два многомерных набора данных размерностью 11×14. Используя методы традиционной и многомерной статистики, была сделана попытка найти закономерности рассеяния локусов и получить интегрированный показатель аллельного разнообразия.

Изменчивость стандартизированных оценок разнообразия/дифференциации в пределах локусов находилась в диапазоне 6-32 %. Непараметрический тест Mann-Whitney-Wilcoxon показал статистически значимые ($p_{Bonf} \leq$ 0,0009) различия медиан локуса 1 (*Eth3*) с локусом 9 (*Bm2113*), локуса 4 (*Tgla126*) с локусами 1, 2, 5, 7, 9, 10, 11 (*Eth3*, *Inra23*, *Tgla122*, *Eth225*, *Bm2113*, *Bm1824*, *Eth10*). PCA выделил две главные компоненты с общей информативностью 95,2 %. Первая учитывала 59,4 % общей

дисперсии, имела более высокие нагрузки по показателями внутрипородного разнообразия и названа «альфа-компонентой». Вторая объясняла 35,8 % общей дисперсии, имела высокие нагрузки по показателям межпородной дифференциации и определена как «бета-компонента». На PCA-ординации локусы 3 и 8 (*Tgla227* и *Tgla53*) формировали группу A, имели высокие величины альфа и ниже среднего бета-компонент. В группу B вошли локусы 5, 7 и 11 (*Tgla122*, *Eth225* и *Eth10*), имевшие выше среднего оценки, как альфа-, так и бета-компонент. Группу C сформировали локусы 2, 9 и 10 (*Inra23*, *Bm2113* и *Bm1824*) с высокими оценками по бета-компоненте. Локусы 1, 4 и 6 (*Eth3*, *Tgla126* и *Sps115*) не относились ни к одной из групп; были определены как «нетипичные» и объединены в условную группу D. Для них были характерны низкие величины, как альфа-, так и бета-компонент. Таким образом, PCA-ординация классифицировала локусы, отразив через реальные взаимосвязи исходных оценок неизвестную до анализа их компонентную структуру (местоположением выделенных групп).

Валидность PCA-ординации подтвердилась расчётами по редуцированным данным (11×7) и методом nMDS. Согласованность ординаций по тесту Прокруста составила $r^2 = 96$ % при $p_{perm} = 0{,}001$. Аналогичная структура локусов была получена кластерным анализом (UPGMA), который определил и бутстрэп-вероятности группировки локусов, близких по латентным переменным: кластера A – 73 %, B – 100 %, C – 73 % и D – 47 %.

Локусы 4 и 6 (*Tgla126* и *Sps115*) имели минимальное сходство с центроидом (S ≈ 40 %); сходство локусов 8 и 10 (*Tgla53* и *Bm1824*) было на уровне 60 %, локусов 2, 3 и 9 (*Inra23*, *Tgla227* и *Bm2113*) – 70-75 %. Наибольшую близость к центроиду имели локусы 1, 5, 7 и 11 (*Eth3*, *Tgla122*, *Eth225* и *Eth10*) – 84-88 %. Среднее абсолютное отклонение оценок показателей разнообразия/дифференциации по четырём локусам с S ≥ 84 % от «истинных» оценок по 11 локусам составило 3,4 % (в пределах статистической ошибки), четырёх локусов с S ≤ 62% – 12,4 % (статистически значимое отклонение). Поэтому локусы *Eth3*, *Tgla122*, *Eth225* и *Eth10* являются, возможно, более подходящими для предварительной декомпозиции аллельного разнообразия и/или каких-либо иных разведочных анализов.

Список литературы

1. Денискова Т. Е., Сермягин А. А., Багиров В. А., Охлопков И. М., Гладырь Е. А., Иванов Р. В., Брем Г., Зиновьева Н. А. Сравнительное исследование информативности STR и SNP маркеров для внутривидовой и межвидовой дифференциации рода *Ovis*. Генетика. 2016;52(1):90-96. DOI: https://doi.org/10.7868/S0016675816010021 EDN: VCPJIJ

2. Сермягин А. А., Белоус А. А., Контэ А. Ф., Филипченко А. А., Ермилов А. Н., Янчуков И. Н., Племяшов К. В., Брем Г., Зиновьева Н. А. Валидация геномного прогноза племенной ценности быков-производителей по признакам молочной продуктивности дочерей на примере популяции черно-пестрого и голштинского скота. Сельскохозяйственная биология. 2017;52(6):1148-1156. DOI: https://doi.org/10.15389/agrobiology.2017.6.1148rus EDN: YLSVCC

3. Смарагдов М. Г., Кудинов А. А. Полногеномная оценка инбридинга у молочного скота. Достижения науки и техники АПК. 2019;33(6):51-53. Режим доступа: https://www.elibrary.ru/item.asp?id=39179451 EDN: UQDHGE

4. Sermyagin A. A., Dotsev A. V., Gladyr E. A., Traspov A. A., Deniskova T. E., Kostyunina O. V., Reyer H., Wimmers K., Barbato M., Paronyan I. A., Plemyashov K. V., Sölkner J., Popov R. G., Brem G., Zinovieva N. A. Whole-genome SNP analysis elucidates the genetic structure of Russian cattle and its relationship with Eurasian taurine breeds. Genetics Selection Evolution. 2018;50:37. DOI: https://doi.org/10.1186/s12711-018-0408-8

5. Volkova V. V., Abdelmanova A. S., Deniskova T. E., Romanenkova O. S., Khozhokov A. A., Ozdemirov A. A., Sermyagin A. A., Zinovieva N. A. Investigation of the Genetic Diversity of Dagestan Mountain Cattle Using STR-Markers. Diversity. 2022;14(7):569. DOI: https://doi.org/10.3390/d14070569

6. Калашникова В. В., Храброва Л. А., Зайцев А. М., Зайцева М. А., Калинкова Л. В. Полиморфизм микросателлитной ДНК у лошадей заводских и локальных пород. Сельскохозяйственная биология. 2011;46(2):41-45.

7. Deniskova T. E., Dotsev A. V., Selionova M. I., Kunz E., Medugorac I., Reyer H., Wimmers K., Barbato M., Traspov A. A., Brem G., Zinovieva N. A. Population structure and genetic diversity of 25 Russian sheep breeds based on whole-genome genotyping. Genetics Selection Evolution. 2018;50:29. DOI: https://doi.org/10.1186/s12711-018-0399-5

8. Харзинова В. Р., Зиновьева Н. А. Паттерн генетического разнообразия у локальных и коммерческих пород свиней на основе анализа микросателлитов. Вавиловский журнал генетики и селекции. 2020;24(7):747-754. DOI: https://doi.org/10.18699/VJ20.669 EDN: BJRYAW

9. Kharzinova V. R., Dotsev A. V., Solovieva A. D., Shimit L. D.-O., Kochkarev A. P., Reyer H., Zinovieva N. A. Genome-Wide SNP Analysis Reveals the Genetic Diversity and Population Structure of the Domestic Reindeer Population (Rangifer tarandus) Inhabiting the Indigenous Tofalar Lands of Southern Siberia. Diversity. 2022;14(11):900. DOI: https://doi.org/10.3390/d14110900

10. 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 EDN: TFRDMN

11. Кузнецов В. М. Методы Нея 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 EDN: DSEMYO

12. 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

13. Jost L. GST and its relatives do not measure differentiation. Molecular Ecology. 2008;17(18):4015-4026. DOI: https://doi.org/10.1111/j.1365-294X.2008.03887.x

14. Chao A., Ma K. H., Hsieh T. C., Chiu C. H. Online Program SpadeR (Species-richnessPrediction And Diversity Estimationin R). Program and User's Guide. 2015. URL: http://chao.stat.nthu.edu.tw/wordpress/software_download/

15. 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

16. 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 EDN: FYQNTE

17. 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 EDN: LGYMFT

18. 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 EDN: LSSUYZ

19. Putman A. I., Carbone I. Challenges in analysis and interpretation of microsatellite data for population genetic studies. Ecology and Evolution. 2014;4(22):4399-4428. DOI: https://doi.org/10.1002/ece3.1305

20. Peakall R., Smouse P. E. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes. 2006;6(1):288-295. DOI: https://doi.org/10.1111/j.1471-8286.2005.01155.x

21. 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

22. Smouse P. E., Whitehead M., Peakall R. An informational diversity framework, illustrated with sexually deceptive orchids in early stages of speciation. Molecular Ecology Resources. 2015;15(6):1375-1384. DOI: https://doi.org/10.1111/1755-0998.12422

23. Hammer Ø., Harper D. A. T., Ryan P. D. PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica. 2001;4(1):1-9.

24. Camúñez L. E. M., Roca C. F., Tornero R. Guía de *KyPlot*: Programa de análisis de datos en contexto científico. Facultat de Física-Universitat de València (UVEG). 2008. 33 p.

25. Jackson D. A. PROTEST: A PROcrustean Randomization TEST of community environment concordance. Ecoscience. 1995;2(3):297-303. DOI: https://doi.org/10.1080/11956860.1995.11682297

26. Peres-Neto P. R., Jackson D. A. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia. 2001;129(2):169-178. DOI: https://doi.org/10.1007/s004420100720

27. Dray S., Chessel D., Thioulouse J. Procrustean co-inertia analysis for the linking of multivariate datasets. Écoscience. 2003;10(1):110-119. DOI: https://doi.org/10.1080/11956860.2003.11682757

28. Kruskal W. H., Wallis W. A. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association. 1952;47(260):583-621. DOI: https://doi.org/10.2307/2280779

29. Ким Дж.-О., Мьюллер Ч. У., Клекка У. Р., Олдендерфер М. С., Блэшфилд Р. К. Факторный, дискриминантный и кластерный анализ. Пер. с англ. Под ред. И. С. Енюкова. М.: «Финансы и статистика», 1989. 215 с.

30. Henderson C. R. Selection index and expected genetic advance. In: «Statistical genetics and plant breeding». Hanson W. D. and Robinson H. F. (eds). NAS-NRS. 1963. Pp.141-163.

References

1. Deniskova T. E., Sermyagin A. A., Bagirov V. A., Okhlopkov I. M., Gladyr E. A., Ivanov R. V., Brem G., Zinoveva N. A. Comparative analysis of the effectiveness of str and snp markers for intraspecific and interspecific differentiation of the genus *Ovis*. *Genetika* = Russian Journal of Genetics. 2016;52(1):90-96. (In Russ.). DOI: https://doi.org/10.7868/S0016675816010021

2. Sermyagin A. A., Belous A. A., Konte A. F., Filipchenko A. A., Ermilov A. N., Yanchukov I. N., Plemyashov K. V., Brem G., Zinoveva N. A. Genomic evaluation of bulls for daughters' milk traits in russian black-and-white and holstein cattle population through the validation procedure. *Sel'skokhozyaystvennaya biologiya* = Agricultural Biology. 2017;52(6):1148-1156. (In Russ.). DOI: https://doi.org/10.15389/agrobiology.2017.6.1148rus

3. Smaragdov M. G., Kudinov A. A. Full genome inbreeding assessment of dairy cattle. *Dostizheniya nauki i tekhniki APK* = Achievements of Science and Technology of AICis. 2019;33(6):51-53. (In Russ.). URL: https://www.elibrary.ru/item.asp?id=39179451

4. Sermyagin A. A., Dotsev A. V., Gladyr E. A., Traspov A. A., Deniskova T. E., Kostyunina O. V., Reyer H., Wimmers K., Barbato M., Paronyan I. A., Plemyashov K. V., Sölkner J., Popov R. G., Brem G., Zinovieva N. A. Whole-genome SNP analysis elucidates the genetic structure of Russian cattle and its relationship with Eurasian taurine breeds. Genetics Selection Evolution. 2018;50:37. DOI: https://doi.org/10.1186/s12711-018-0408-8

5. Volkova V. V., Abdelmanova A. S., Deniskova T. E., Romanenkova O. S., Khozhokov A. A., Ozdemirov A. A., Sermyagin A. A., Zinovieva N. A. Investigation of the Genetic Diversity of Dagestan Mountain Cattle Using STR-Markers. Diversity. 2022;14(7):569. DOI: https://doi.org/10.3390/d14070569

6. Kalashnikova V. V., Khrabrova L. A., Zaytsev A. M., Zaytseva M. A., Kalinkova L. V. Polymorphism of microsatellite dna in horses of stud and local breeds. *Sel'skokhozyaystvennaya biologiya* = Agricultural Biology. 2011;46(2):41-45. (In Russ.).

7. Deniskova T. E., Dotsev A. V., Selionova M. I., Kunz E., Medugorac I., Reyer H., Wimmers K., Barbato M., Traspov A. A., Brem G., Zinovieva N. A. Population structure and genetic diversity of 25 Russian sheep breeds based on whole-genome genotyping. Genetics Selection Evolution. 2018;50:29. DOI: https://doi.org/10.1186/s12711-018-0399-5

8. Kharzinova V. R., Zinovieva N. A. The pattern of genetic diversity of different breeds of pigs based on microsatellite analysis. *Vavilovskiy zhurnal genetiki i selektsii* = Vavilov Journal of Genetics and Breeding. 2020;24(7):747-754. (In Russ.). DOI: https://doi.org/10.18699/VJ20.669

9. Kharzinova V. R., Dotsev A. V., Solovieva A. D., Shimit L. D.-O., Kochkarev A. P., Reyer H., Zinovieva N. A. Genome-Wide SNP Analysis Reveals the Genetic Diversity and Population Structure of the Domestic Reindeer Population (Rangifer tarandus) Inhabiting the Indigenous Tofalar Lands of Southern Siberia. Diversity. 2022;14(11):900. DOI: https://doi.org/10.3390/d14110900

10. Kuznetsov V. M. 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

11. Kuznetsov V. M. 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

12. 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

13. Jost L. GST and its relatives do not measure differentiation. Molecular Ecology. 2008;17(18):4015-4026.
DOI: https://doi.org/10.1111/j.1365-294X.2008.03887.x

14. Chao A., Ma K. H., Hsieh T. C., Chiu C. H. Online Program SpadeR (Species-richnessPrediction And Diversity Estimationin R). Program and User's Guide. 2015. URL: http://chao.stat.nthu.edu.tw/wordpress/software_download/

15. 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

16. Kuznetsov V. M. 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

17. Kuznetsov V. M. 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

18. 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

19. Putman A. I., Carbone I. Challenges in analysis and interpretation of microsatellite data for population genetic studies. Ecology and Evolution. 2014;4(22):4399-4428. DOI: https://doi.org/10.1002/ece3.1305

20. Peakall R., Smouse P. E. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes. 2006;6(1):288-295. DOI: https://doi.org/10.1111/j.1471-8286.2005.01155.x

21. 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

22. Smouse P. E., Whitehead M., Peakall R. An informational diversity framework, illustrated with sexually deceptive orchids in early stages of speciation. Molecular Ecology Resources. 2015;15(6):1375-1384. DOI: https://doi.org/10.1111/1755-0998.12422

23. Hammer Ø., Harper D. A. T., Ryan P. D. PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica. 2001;4(1):1-9.

24. Camúñez L. E. M., Roca C. F., Tornero R. Guía de *KyPlot*: Programa de análisis de datos en contexto científico. Facultat de Física-Universitat de València (UVEG). 2008. 33 p.

25. Jackson D. A. PROTEST: A PROcrustean Randomization TEST of community environment concordance. Ecoscience. 1995;2(3):297-303. DOI: https://doi.org/10.1080/11956860.1995.11682297

26. Peres-Neto P. R., Jackson D. A. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia. 2001;129(2):169-178. DOI: https://doi.org/10.1007/s004420100720

27. Dray S., Chessel D., Thioulouse J. Procrustean co-inertia analysis for the linking of multivariate datasets. Écoscience. 2003;10(1):110-119. DOI: https://doi.org/10.1080/11956860.2003.11682757

28. Kruskal W. H., Wallis W. A. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association. 1952;47(260):583-621. DOI: https://doi.org/10.2307/2280779

29. Kim Dzh.-O., M'yuller Ch. U., Klekka U. R., Oldenderfer M. S., Bleshfild R. K. Factorial, discriminant and cluster analysis. *Per. s angl. Pod red. I. S. Enyukova*. Moscow: *«Finansy i statistika»*, 1989. 215 p.

30. Henderson C. R. Selection index and expected genetic advance. In: «Statistical genetics and plant breeding». Hanson W. D. and Robinson H. F. (eds). NAS-NRS. 1963. Pp.141-163.

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

✉ **Vasiliy 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