

<https://doi.org/10.30766/2072-9081.2025.26.4.852-871>

УДК 636.082:575.174.015.3

Информационный контент полиморфизма STR-локусов в выборках быков-производителей трёх пород

© 2025. В. М. Кузнецов✉

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

Данные генотипирования по 11 микросателлитным локусам (STR) производителей джерсейской (JER; $n = 10$), красной скандинавской (RED; $n = 29$) и голштинской (HOL; $n = 45$) пород использовали для расчёта информационного контента полиморфизма (PIC). По JER-выборке оценки PIC локусов были в диапазоне [0,222; 0,680], по RED-выборке – [0,448; 0,802], по HOL-выборке – [0,466; 0,825]; средние соответственно 0,470, 0,650 и 0,682. Различия среднего по JER-выборке со средними по RED- и HOL-выборкам были статистически значимыми ($p_{\text{value}} < 0,005$). Высокоинформативных локусов ($\text{PIC} > 0,6$) в JER-, RED- и HOL-выборках было 36,4, 63,6 и 72,7 % соответственно. В 2D-проекции анализа соответствия породных выборок с четырьмя категориями – первая размерность объясняла 94,5 % извлеченной инверсии, вторая – 4,5 %. На ординации проявилась близость RED- и HOL-выборок и их контраст с JER-выборкой. Также имели место близость 2-ой ($\text{PIC} = 0,4-0,6$) и 3-ей ($\text{PIC} = 0,61-0,8$) категорий, их контраст с 4-ой категорией ($\text{PIC} > 0,8$) и более значительный – с 1-ой категорией ($\text{PIC} < 0,4$). Для идентификации животных с ошибкой 0,0001 в JER-выборке было достаточно пять, в RED-выборке – четыре, в HOL-выборке – три локуса с высокими показателями PIC. При проверке происхождения, когда генотипы обоих родителей известны, вероятность исключения 99,9 % достигалась в HOL-выборке по 8 локусам, в RED-выборке – по 10, в JER-выборке требовалось более 11 локусов. В случае, когда известен генотип одного родителя, все 11 локусов в JER-, RED- и HOL-выборках могли обеспечить вероятность исключения 88,2, 98,3 и 99,1 % соответственно. Показатели индивидуальной гетерозиготности одних и тех же производителей, рассчитанные по высокоинформативным и низкоинформативным локусам, были статистически независимыми ($r^2 = 0,07$). Оценки индексов фиксации (G_{ST} , $G'_{\text{ST}(N)}$), их модификаций ($G'_{\text{ST}(H)}$, G''_{ST}) и межпородной дифференциации (D_{est} , D') RED- и HOL-выборок были: по 11 локусам – 0,056 и 0,105, 0,331 и 0,366, 0,292 и 0,343; по пяти низкоинформативным локусам (PIC_{min}) – 0,07 и 0,13, 0,292 и 0,338, 0,238 и 0,269; по пяти высокоинформативным локусам (PIC_{max}) – 0,034 и 0,066, 0,319 и 0,342, 0,295 и 0,355 соответственно. При планировании широкомасштабных популяционно-генетических исследований выбор высокоинформативных маркеров ДНК, по крайней мере, не снизит точность генетических оценок и тестов, но сократит затраты на генотипирование и анализы меньшего числа локусов.

Ключевые слова: породы КРС, STR-маркеры, генетическое разнообразие, PIC, ANOVA, сеть, вероятность идентичности, вероятность исключения, индивидуальная гетерозиготность, анализ соответствия, меры дифференциации

Благодарности: работа выполнена при поддержке Минобрнауки РФ в рамках Государственного задания ФГБНУ «Федеральный аграрный научный центр Северо-Востока имени Н. В. Рудницкого» (№ гос. регистрации 1021060407726-4).

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

Конфликт интересов: автор заявил об отсутствии конфликта интересов.

Для цитирования: Кузнецов В. М. Информационный контент полиморфизма STR-локусов в выборках быков-производителей трёх пород. Аграрная наука Евро-Северо-Востока. 2025;26(4):852–871.

DOI: <https://doi.org/10.30766/2072-9081.2025.26.4.852-871>

Поступила: 09.02.2025

Принята к публикации: 12.08.2025

Опубликована онлайн: 29.08.2025

The polymorphism information content of the STR loci in the samples of breeding bulls of three breeds

© 2025. Vasily M. Kuznetsov✉

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

Genotyping data for 11 microsatellite loci (STR) of the Jersey (JER, $n = 10$), Red Scandinavian (RED, $n = 29$) and Holstein (HOL, $n = 45$) breeds were used to calculate polymorphism information content (PIC). For the JER-sample, the estimates of PIC loci were in the range of [0,222; 0,680], for the RED-sample – [0,448; 0,802], for the HOL-sample – [0,466; 0,825]; averages were 0,470, 0,650 and 0,682, respectively. The differences between the average of the JER-sample and the average of the RED- and HOL-samples were statistically significant ($p_{\text{value}} < 0,005$). Highly informative loci ($\text{PIC} > 0,6$) in the JER-, RED- and HOL-samples were 36,4, 63,6 and 72,7 %, respectively. In the 2D projection of the analysis of the correspondence of bull samples with four PIC categories, the first dimension explained 94,5 % of the extracted inversion,

the second – 4.5 %. The ordination revealed the proximity of the RED- and HOL-samples and their contrast with the JER-sample. There was also a proximity of 2 ($PIC = 0.4-0.6$) and 3 ($PIC = 0.61-0.8$) categories, their contrast with category 4 ($PIC > 0.8$) and a stronger contrast with category 1 ($PIC < 0.4$). To identify animals with an error of 0.0001, five loci with high PIC scores were sufficient in the JER-sample, four in the RED-sample, and three in the HOL-sample. When verifying the origin, when the genotypes of both parents are known, the 99.9 % probability exclusion was achieved in the HOL-sample at 8 loci, in the RED-sample at 10, and in the JER-sample more than 11 loci were required. In the case where the genotype of one parent is known, all 11 loci in the JER-, RED- and HOL-samples could provide probability exclusion of 88.2, 98.3, and 99.1 %, respectively. The indicators of individual heterozygosity of the same bulls, calculated from highly and low-informative loci, were statistically independent ($r^2 = 0.07$). Estimates of fixation indices (G_{ST} , $G'_{ST(N)}$), their modifications ($G'_{ST(H)}$, G''_{ST}) and interbreed differentiation (D_{est} , D') the RED- and HOL-samples were: for 11 loci – 0.056 and 0.105, 0.331 and 0.366, 0.292 and 0.343, respectively; for five low-information loci (PIC_{min}) – 0.07 and 0.13, 0.292 and 0.338, 0.238 and 0.269; for five high-information loci (PIC_{max}) – 0.034 and 0.066, 0.319 and 0.342, 0.295 and 0.355. When planning large-scale population-genetic studies, the choice of highly informative DNA markers will at least not reduce the accuracy of genetic assessments and tests, but will cut back the cost of genotyping and analysis of fewer loci.

Keywords: cattle breeds, STR markers, genetic diversity, Polymorphism Information Content – PIC, ANOVA, network, probability of identity, probability of exclusion, individual heterozygosity, correspondence analysis, differentiation measures

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. 1021060407726-4).

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 citations: Kuznetsov V. M. The polymorphism information content of the STR loci in the samples of breeding bulls of three breeds. *Agrarnaya nauka Evro-Severo-Vostoka* = Agricultural Science Euro-North-East. 2025;26(4):852–871. (In Russ.). DOI: <https://doi.org/10.30766/2072-9081.2025.26.4.852-871>

Received: 09.02.2025

Accepted for publication: 12.08.2025

Published online: 29.08.2025

Прогресс в картировании генов обеспечил генетиков и селекционеров большим числом микросателлитных маркеров (Simple Tandem Repeats, STR), ассоциированных с локусами, для которых характерна мультиаллельность. Для поиска информативных локусов, определения необходимого их числа и должного уровня полиморфизма используют два показателя: H_e – ожидаемая гетерозиготность и PIC (Polymorphism Information Content) – информационный контент полиморфизма. Оба показателя варьируют в диапазоне [0; 1] или [0; 100 %].

Ожидаемая гетерозиготность, H_e , является фундаментальной мерой генетической изменчивости в популяции, показывающая долю гетерозиготных генотипов, которая может быть при равновесии Харди-Вайнберга с теми же аллельными частотами, что и в реальной популяции [1]. Для локуса с n аллелями и частотой i -го аллели p_i

$$H_e = 1 - \sum_{i=1}^n p_i^2,$$

где $\sum_{i=1}^n p_i^2$ – это частота всех гомозиготных генотипов; считается, что гомозиготы не несут информации. М. Nei [1] назвал H_e – генным разнообразием (вероятность того, что две аллели, случайным образом извлечённые из популяции, различны или вероятность того, что любой случайно выбранный индивид

из популяции будет гетерозиготным по любым двум аллелям в маркерном локусе).

Показатель PIC предложили D. Botstein с соавторами [2] для исследования сцепления редкого доминантного заболевания, когда поражён один из родителей, который имеет гетерозиготный генотип и аллель предрасположенности к заболеванию¹. В этом контексте термин «информационный» означал – пригодный для анализа сцепления того, что маркер находится рядом с геном или другим представляющим интерес маркером. Для локуса с частотами i -го и j -го аллелей p_i и p_j :

$$PIC = 1 - (\sum_{i=1}^n p_i^2) - \sum_{i=1}^{n-1} \sum_{j=i+1}^n 2p_i^2 p_j^2.$$

Третий элемент в правой части – это вероятность того, что оба родителя и потомок будут иметь одинаковые гетерозиготные генотипы. В сущности, PIC – это модифицированный показатель H_e с поправкой на то, что индивид при анализе сцепления не внесёт ошибки в результаты [3].

Х. Го, Р. К. Элстон (X. Guo, R. C. Elston) [4] обобщили определение PIC таким образом, чтобы оно не зависело от типа наследования изучаемого признака. Для этого они ввели понятие *информативность передачи* (transmission informativeness), под которым понимали

¹Предполагали: (а) в общей популяции частоты аллелей в маркерном локусе не зависят от частот аллелей в локусе заболевания и (б) пара аллелей встречается с частотой, равной произведению двух частот.

условную вероятность² передачи родителем одного из двух маркерных аллелей (при известных маркерных генотипах потомка и обоих родителей). Информативность передачи маркера для триады «отец-мать-потомок» в нуклеарной семье равна 0 или 1. Обобщённый показатель *PIC* – это информативность передачи, усреднённая по всем возможным триадам, т. е. безусловная вероятность того, какой маркерный аллель передал родитель. Чтобы триада была информативной для родителя, этот родитель должен быть гетерозиготным, а родители и потомок не должны иметь один и тот же гетерозиготный генотип. Х. Го, Р. К. Элстон (X. Guo, R. C. Elston) пришли к выводу, что Botstein's формула *PIC* является общей для оценивания информативности маркерного локуса, независимо от типа наследования сцеплённого признака.

Как мера полиморфизма, *PIC* зависит только от числа аллелей и частоты каждой аллели в маркерном локусе. Локусы с большим числом аллелей или/и равными аллельными частотами имеют более высокие показатели *PIC*. Они обладают большей способностью различать тестируемые образцы (высокой дискриминационной мощностью).

В литературе встречаются другие формулы расчёта *PIC*:

$$PIC = 1 - S_2 - S_2^2 + S_4, \dots \text{ где } S_k = \sum_{i=1}^n p_i^k \quad [5]$$

$$PIC = 2 \sum_{i=1}^n \sum_{j=1}^n p_i \times p_j (1 - p_i \times p_j) \quad [6]$$

$$PIC = 1 - \sum_{i=1}^n p_i^2 - \left(\sum_{i=1}^n p_i^2 \right)^2 + \sum_{i=1}^n p_i^4 \quad [7]$$

$$I_2 = 2 \sum_{u=1}^{n-1} \sum_{v=u+1}^n p_u p_v - 2 \sum_{u=1}^{n-1} \sum_{v=u+1}^n p_u^2 p_v^2 \quad [8]$$

где I_2 – прогноз *PIC*; n – число аллелей; p_u и p_v – частоты маркерных аллелей u и v гетерозиготного родителя.

По Botstein's *PIC* представляет вероятность того, что данный потомок родителя, который имеет редкую аллель в индексном локусе (предрасположенность к заболеванию), позволит определить родительский генотип маркерного локуса. По [9] *PIC* – это вероятность получения потомства, для которого аллельное происхождение одного родителя может определяться по случайному спариванию. По [4] – это вероятность того, что маркерный генотип потомка позволит при отсут-

ствии кроссинговера определить, какой из двух маркерных аллелей поражённого родителя он получил, т. е. *PIC* является вероятностью получения аллели от отца или от матери. Если маркерные аллели, которые были переданы отцом и матерью, известны, или если можно определить от какого родителя потомок получил каждую аллель, то можно установить пропорцию аллелей, идентичных по происхождению (Identical by Descent, IBD). Чем выше показатель *PIC*, тем более информативен маркер и, следовательно, выше вероятность определения доли IBD-аллелей. По [7] *PIC* маркера определяется как ожидаемая доля информативного потомства у данного типа родословной. Потомство является информативным, если по его генотипу можно сделать вывод о маркерной аллели, связанной с аллелью заболевания, и, следовательно, совместно наследуемых в последующих поколениях. Полиморфизм на уровне выборки может быть предсказан по ожидаемой частоте информативных потомков в выборке (*PIC*), а полиморфизм на уровне семьи – по частоте информативных потомков на одного гетерозиготного родителя, который называют *информационным контентом семьи* (Family Information Content, *FIC*) [8]. *PIC* объединяет гетерозиготность родителей с долей потомства, для которого известно наследование маркера. Наследование неизвестно, если потомок имеет тот же маркерный генотип, что и оба родителя, или как у одного из родителей, а другой родитель неизвестен [10]. По [6] *PIC* – это способность маркера дифференцировать особей в популяции; по [11] *PIC* отражает способность специфического маркера обнаруживать полиморфизм в изучаемых образцах; по [12] *PIC* – это мера количества информации, которая может быть извлечена из генетического маркера.

Цель исследования – дать характеристику выборок быков-производителей трёх пород по информационному контенту полиморфизма STR-локусов с последующим генетико-статистическим анализом *PIC*-оценок. В частности, ставились задачи оценить взаимосвязи *PIC*-показателей с иными характеристиками разнообразия, исследовать межпородные различия, изучить вероятности идентификации генотипов и способность STR-локусов выявлять несоответствия в записях родословных.

²Условная вероятность – это вероятность наступления события А при условии, что событие В произошло. Безусловная вероятность – это вероятность события независимо от предшествующего или будущего возникновения других событий (просто вероятность наступления события).

Научная новизна – проведён сетевой анализ локусов при различных ограничениях по индексу сходства *PIC*-профилей; изучена сопряжённость породных выборок с *PIC*-категориями; исследовано влияние уровня *PIC* локусов на оценки индивидуальной гетерозиготности производителей и относительной генетической дифференциации выборок.

Материал и методы. Использовали данные генотипирования по 11 STR-локусам (*Eth3*, *inra23*, *Tgla227*, *Tgla126*, *Tgla122*, *Sps115*, *Eth225*, *Tgla53*, *Bm2113*, *Bm1824*, *Eth10*) 84 быков-производителей трёх пород³: джерсейской породы (*JER*; $n = 10$), красной скандинавской (*Viking Red*, *RED*; $n = 29$) и голштинской (*HOL*; $n = 45$)⁴. Кроме показателей H_e и *PIC* (по [2]) рассчитывали: энтропийный аналог эффективного числа аллелей на локус –

$s_{n_e} = \exp(sH)$, где sH – индекс Шеннона [13, 14]⁵; вероятности идентичности животных и исключения предполагаемого родителя(-ей); индивидуальную гетерозиготность; статистики относительной генетической дифференциации выборок: индексы фиксации (G_{ST} [16], $G'_{ST(N)}$ [1, 17]), модифицированные по [17] индексы дифференциации (индексы фиксации с поправкой на внутривыборочную гетерозиготность – $G'_{ST(H)}$ и G''_{ST}) и «подлинные» индексы дифференциации (D_{est} [18], D' [14, 19]; см. также [20, 21]). Использованы программы GenAlEx 6.502 [22, 23], PAST3 [24], STATGRAPHICS®CenturionXVI [25].

Результаты и их обсуждение. В таблице 1 даны оценки и описательная статистика показателей разнообразия s_{n_e} , H_e и *PIC* по локусам в пределах *JER*-, *RED*- и *HOL*-выборок.

Таблица 1 – Оценки s_{n_e} , H_e и *PIC* по локусам и статистика по выборкам /
Table 1 – Estimates s_{n_e} , H_e and *PIC* by loci and statistics by samples

Локус / Loci	s_{n_e}			H_e			<i>PIC</i>		
	<i>JER</i>	<i>RED</i>	<i>HOL</i>	<i>JER</i>	<i>RED</i>	<i>HOL</i>	<i>JER</i>	<i>RED</i>	<i>HOL</i>
<i>Eth3</i> (1)	1,75	4,16	3,22	0,375	0,701	0,638	0,304	0,661	0,569
<i>inra23</i> (2)	2,75	3,34	4,34	0,615	0,574	0,747	0,533	0,541	0,703
<i>Tgla227</i> (3)	3,89	6,49	6,86	0,710	0,823	0,825	0,658	0,802	0,804
<i>Tgla126</i> (4)	3,18	3,48	2,52	0,585	0,645	0,543	0,543	0,595	0,466
<i>Tgla122</i> (5)	3,58	3,87	6,23	0,690	0,659	0,815	0,639	0,618	0,792
<i>Sps115</i> (6)	1,68	2,67	3,35	0,265	0,474	0,614	0,246	0,448	0,570
<i>Eth225</i> (7)	4,12	5,69	3,91	0,725	0,791	0,681	0,680	0,761	0,627
<i>Tgla53</i> (8)	3,34	5,37	7,92	0,675	0,737	0,843	0,611	0,710	0,825
<i>Bm2113</i> (9)	1,53	5,06	5,01	0,255	0,766	0,776	0,222	0,735	0,739
<i>Bm1824</i> (10)	1,75	2,96	3,73	0,375	0,549	0,685	0,305	0,513	0,634
<i>Eth10</i> (11)	2,71	5,61	5,75	0,595	0,796	0,797	0,430	0,768	0,770
Min	1,53	2,67	2,52	0,255	0,474	0,543	0,222	0,448	0,466
Max	4,12	6,49	7,92	0,725	0,823	0,843	0,680	0,802	0,825
Range	2,59	3,82	5,40	0,470	0,349	0,300	0,458	0,354	0,359
CV, %	34,5	28,8	35,4	33,8	16,7	13,6	37,2	17,9	17,1
Mean	2,75	4,43	4,80	0,533	0,683	0,724	0,470	0,650	0,682
SE	0,287	0,384	0,513	0,054	0,034	0,030	0,053	0,035	0,035
95 % CIL	2,11	3,57	3,66	0,412	0,607	0,658	0,353	0,572	0,603
95 % CIU	3,39	5,28	5,95	0,654	0,760	0,790	0,587	0,728	0,760

Примечание: Range = Max – Min; CV – коэффициент изменчивости; Mean – среднее; SE – стандартная ошибка; 95 % CIL/95 % CIU – 95%-ный доверительный интервал (нижняя/верхняя границы) /

Notes: CV – coefficient of variability; SE – standard error; 95 % CIL/95 % CIU – 95 % confidence interval (lower/upper bounds).

³База генетических данных быков-производителей. Микросателлиты. [Электронный ресурс].

URL: <http://www.vniiplem.ru/rus/files/Database/DNK/mikrosatellity.pdf> (дата обращения: 09.12.2018).

⁴Из 33 тестов согласия распределений генотипов с равновесием Харди-Вайнберга (ПХВ) в двух случаях имело место статистически значимое отклонение от ПХВ: по локусу *Eth10* в *JER*-выборке и по локусу *Eth225* в *HOL*-выборке.

⁵Наблюдаемое число аллелей не является подходящим показателем полиморфизма маркеров [8]. Эффективное число аллелей на локус, рассчитываемое традиционным методом ($n_e = 1 / \sum_{i=1}^n p_i^2$), занижает оценку аллельного разнообразия, т. к. более распространённым аллелям придаётся больший «вес». Статистика s_{n_e} лишена этого недостатка [15].

Оценки в пределах выборок были проверены на нормальность распределения. Одномерный тест Жарке-Бера (JB)⁶ показал значения достигнутой статистической значимости (p_{value}) JB-критерия для JER-выборки в диапазоне [0,517; 0,610], для RED-выборки – [0,654; 0,670], для HOL-выборки – [0,559; 0,682]. Все значения p_{value} были больше принятого уровня критерия значимости ($\alpha = 0,05$) – данные соответствовали совокупности допущений, позволяющих применять параметрические методы статистики.

Описательная статистика дана в нижней части таблицы 1. В JER-выборке коэффициенты изменчивости (CV) оценок s_{n_e, H_e} и PIC были высокими и, примерно, равными (34–37 %). В RED- и HOL-выборках высокую изменчивость имели только s_{n_e} – оценки (28,8 и 35,4 % соответственно); оценки H_e и PIC локусов были более однородными (CV = 14–18 %).

Таблица 2 – Результаты ANOVA по PIC /
Table 2 – Results ANOVA by PIC

Источник / Source	SS	η^2 , %	DF	MS	F-test	p_{value}
Порода / Breed	0,2868	33,2	2	0,1434	12,19	0,0003
Локус / Loci	0,3420	39,6	10	0,0342	2,91	0,0202
Остаток / Residual (e)	0,2352	27,2	20	0,0118	-	-
Сумма / Total	0,8640	100,0	32	-	-	-

Примечания: SS – сумма квадратов (Type III); η^2 – вклад фактора; DF – степени свободы; MS – средний квадрат /
Notes: SS – Sum of Square (Type III); η^2 – the contribution of the factor; DF – Degrees of Freedom; MS – Mean of Square

Коэффициент детерминации статистической модели, равный 72,8 % ($p_{value} = 0,0016$), свидетельствовал о сильной (0,85) статистически значимой связи между PIC и факторами модели. Частный коэффициент детерминации по фактору «порода» 33,2 % соответствовал корреляции 0,58 («заметная»).

Среднее PIC локусов в JER-выборке статистически значимо отличалось от средних RED- и HOL-выборок (табл. 3). Различие между RED- и HOL-выборками было статистически незначимым⁸.

Относительно размаха оценок (Range) выборки распределялись: по s_{n_e} – HOL >> RED >> JER, по H_e – JER >> RED > HOL, по PIC – JER >> HOL $\approx$ RED (>> – существенно больше, $\approx$ – почти равно). Минимальные значения s_{n_e, H_e} и PIC в JER-выборке были по локусу *Bm2113*⁷, в RED-выборке – по локусу *Sps115*, в HOL-выборке – по локусу *Tgla126*; максимальные – по локусам *Eth225*, *Tgla227* и *Tgla53* соответственно.

Ожидаемая гетерозиготность локусов была выше PIC -оценок. Парные тесты Стьюдента показали в JER-выборке различие $0,063 \pm 0,0116$ при $p_{value} = 0,0003$ с 95 % CI [0,037; 0,089]. В RED-выборке – $0,033 \pm 0,0025$, $p_{value} < 0,0001$ и 95 % CI [0,027; 0,039], в HOL-выборке – $0,042 \pm 0,0058$, $p_{value} < 0,001$ и 95 % CI [0,029; 0,055] соответственно.

PIC -оценки локусов были обработаны по двухфакторной модели фиксированного типа (ANOVA): $PIC = \mu + Breed + Loci + e$ (табл. 2).

Таблица 3 – Мульти-сравнение PIC средних /
Table 3 – Multi-comparison of PIC averages

Breeds	JER-RED	JER-HOL	RED-HOL
Differences	-0,180	-0,212	-0,032
p_{value}	0,0047	0,0024	0,3829
95 % CIL	-0,297	-0,329	-0,149
95 % CIU	-0,063	-0,095	0,085

⁶Тест Jarque-Bera (Жарке-Бера, Харке-Бера) – статистический тест, проверяющий ошибки наблюдений на нормальность посредством сверки их третьего момента (асимметрия) и четвёртого момента (экспесс) с моментами нормального распределения. Если p_{value} больше принятого критического α -уровня, то нулевая гипотеза (H_0) о том, что данные получены из нормально распределённой совокупности, не может быть отвергнута.

⁷В JER-выборке слабо информативными ($PIC < 0,25$) были локусы *Bm2113* и *Sps115*.

⁸Сходный результат был по s_{n_e} и H_e переменным.

Корреляции Пирсона между показателями в пределах выборок были высокие и статистически значимые ($p_{value} < 0,0001$), именно: $S_{n_e} \times H_e - 0,94-0,96$, $S_{n_e} \times PIC - 0,95-0,98$, $H_e \times PIC - 0,980-0,998$. По объединённым данным 0,89, 0,92 и 0,99. Регрессия PIC на S_{n_e} в JER-выборке составила $0,180 \pm 0,012$, в RED-выборке была в 2 раза ниже ($0,090 \pm 0,007$), а в HOL-выборке – в 3 раза ($0,060 \pm 0,008$).

На рисунке 1 (слева) дана **лепестковая диаграмма** (radar/spiderplot) профилей выборок по PIC -оценкам локусов. Каждый локус имеет собственную ось координат. Линии соединили все значения PIC локусов (координаты) в каждой выборке и сформировали контуры. Конфигурации контуров и их площади дают наглядное

представление о структуре и уровне полиморфизма выборок. **Диаграмма удобна для сравнения** локусов в выборке, локуса в разных выборках и выборок в целом. Площадь области, ограниченная контуром JER-выборки, заметно меньше, а её конфигурация существенно отличалась от таковых RED- и HOL-выборок. Между последними эти различия были менее выражены. Тем не менее, корреляции Пирсона и Спирмена PIC -оценок локусов между выборками были статистически незначимые ($p_{value} > 0,1$).

На рисунке 1 (справа) рассеяние локусов по PIC -оценкам в пространстве JER×RED×HOL пород. Выделились две группы, близкие по PIC локусов: низкополиморфные (6, 10, 1) и высокополиморфные (2, 5, 8, 3 и 11).

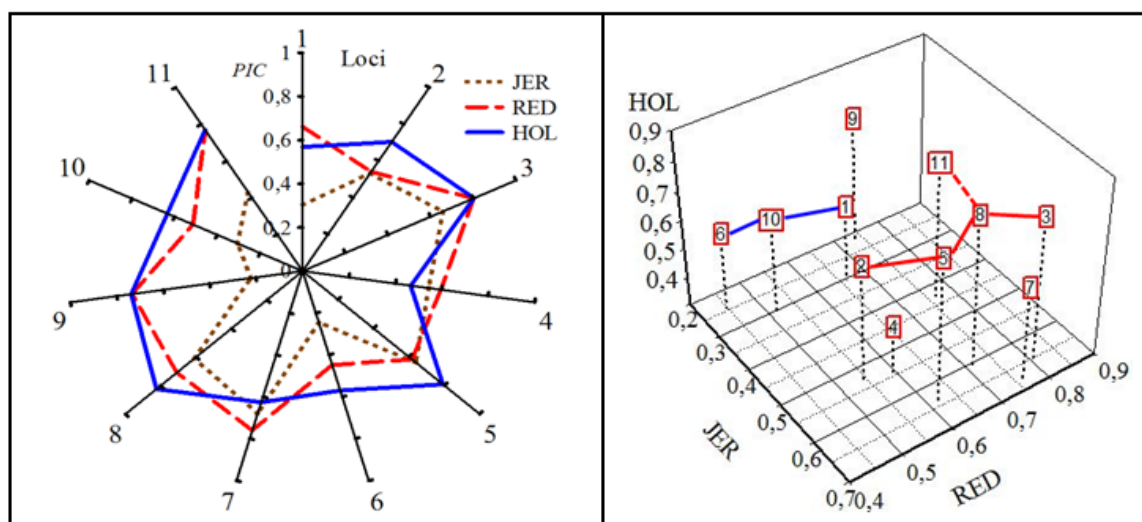


Рис. 1. 2D лепестковая диаграмма PIC -профилей выборок (слева) и рассеяние PIC -оценок локусов в 3D пространстве (справа) /

Fig. 1. 2D spider plot of PIC -profiles of samples (left) and scattering of PIC -estimates of loci in 3D space (right)

Построение сети (networkplot). По PIC -оценкам из таблицы 1 проведён сетевой анализ локусов, который базируется на теории графов. Граф – это совокупность объектов, некоторые пары которых в некотором смысле «связаны». По форме граф представляет сеть из узлов (вершин) и рёбер (граней)⁹. Преимущество сетевого анализа в том, что он придаёт значение отношениям между объектами, а не рассматривает их как изолированные.

Для компоновки сети использовался *Fruchterman-Reingold* алгоритм (FR)¹⁰, который предполагает создание графа с рёбрами одинаковой длины и минимально возможным числом пересечений. На рисунке 2 узлы в сети – это локусы, которые соединены линиями (рёбра), указывающими на наличие связи между ними. Размер (степень) узла пропорционален числу смежных ему рёбер. Узел/локус с высокой степенью является более «влиятельным»¹¹. Этот показатель позволяет дифференцировать узлы на «центральные» и «периферийные».

⁹Граф, в котором рёбра не имеют определённого направления, есть *ненаправленный*.

¹⁰Fruchterman T. M. J., Reingold E. M. Graph drawing by force-directed placement. Software: PracticeandExperience. 1991;21(11):1129–1164. URL: <http://dx.doi.org/10.1002/spe.4380211102>

¹¹Узлы более значимы, если они связаны с узлами большего размера, и снижают свою значимость при контакте с узлами меньшего размера.

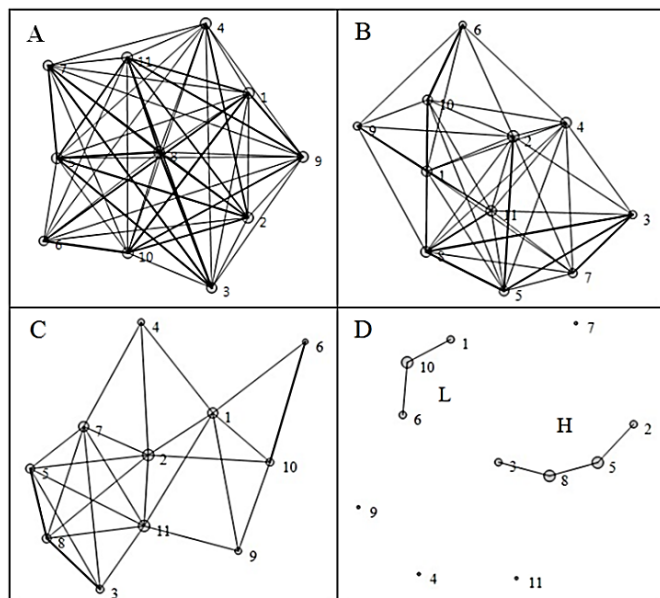


Рис. 2. Компоновка сети по евклидовому сходству (S) PIC -профилей 11 STR-локусов /
Fig. 2. Network layout based on the Euclidean similarity (S) of PIC profiles of 11 STR loci (A – $S \geq 10\%$; B – $S \geq 30\%$; C – $S \geq 50\%$; D – $S \geq 70\%$)

Ребро представляет *отношение подобия* пары локусов, которое измерялось *евклидовым сходством* (S) их профилей (PIC по JER-, RED- и HOL-выборкам). Локусы с похожими профилями (большим индексом) находятся ближе друг к другу. Жирные линии указывают на более высокое сходство. Размер узла и толщина рёбер характеризуют «экспрессию» локуса. Длина ребра определяется FR-алгоритмом только для целей визуализации и не несёт биологической информации.

Число узлов (n) является показателем порядка графа, число рёбер (m) – его размером. В графе порядка n максимальная степень каждого узла равна $n - 1$, а максимальное число рёбер – $n(n - 1) / 2$. Для 11 локусов максимальная степень каждого равна 10, а размер графа – 55.

На рисунке 2-A дан мультиграф с ограничением по индексу сходства $S \geq 10\%$ ($m = 52$). Узел 8 (центральный) и два узла наружного круга 4 и 9 имели степень 10. Все узлы внутреннего круга (1, 2, 10, 11 и 5) имели степень 9. Выделялись циклы узлов с повышенным сходством (1-6-10-1; 2-5-11-2) и цепи (3-5-2-10; 7-8-2-4). Вместе с тем, диапазон значений индекса сходства был очень широкий ($S \geq 10\%$), поэтому сеть запутанная и трудная для интерпретации¹².

Сокращение числа рёбер достигалось повышением требования к минимальному значению индекса сходства. На рисунке 2-B, C, D показана компоновка сетей с нижним порогом

сходства смежных локусов 30, 50 и 70 %¹³. В сети при условии $S \geq 30\%$ (рис. 2-B) общее число рёбер-связей сократилось до 49 со средним индексом ($\bar{S}$) 50,6 %. Локус 2 имел 10 рёбер, локусы 1, 4, 11 – по 9 рёбер. Повышенное сходство профилей имели локусы в цикле 8-3-5-8 ($\bar{S} = 75,7\%$), в цепях 1-10-6 ($\bar{S} = 74,9\%$), 7-3-5-2 ($\bar{S} = 68,0\%$) и 5-8-11 ($\bar{S} = 72,5\%$).

В сети с $S \geq 50\%$ (рис. 2-C) число межлокусных связей сократилось до 26 ($\bar{S} = 61,6\%$). Локусы 2 и 11 имели по 7 рёбер, локусы 1 и 7 – по 6, локусы 5 и 8 – по 5 рёбер. Повышенным сходством характеризовались локусы в цепях 3-8-5 ($\bar{S} = 73,6\%$) и 6-10 ($\bar{S} = 79,5\%$).

При условии $S \geq 70\%$ (рис. 2-D) осталось четыре тривиальных или нулевых графа (локусы 4, 11, 7 и 9), которые не имели связей со сходством более 70 %, и два цепных графа. Граф L составил цепь смежных локусов 6-10-1, граф H – цепь из локусов 3-8-5-2. В этих цепях «влиятельными» были локусы 5, 8 и 10. Каждый из них имел сходство профилей со смежными локусами более 70 %. Локусы 6 и 1, 3 и 2 были «висячие», т. к. имели связи только с одним локусом.

По каждому локусу L- и H-графов были рассчитаны PIC -средние [по породам]. Для локусов L-графа оценки составили 0,421, 0,484 и 0,511 ($\overline{PIC} = 0,472$), для локусов H-графа – 0,592, 0,683, 0,715 и 0,755 ($\overline{PIC} = 0,686$). Следовательно,

¹²В сетевом анализе такой граф называют «hairball» – комок шерсти.

¹³Размещение не строго определено – разные запуски программы приводят к различным конфигурациям сети.

сетевой анализ выявил кластеры локусов с низким и высоким информационным контентом полиморфизма. Отметим похожую группировку локусов по *PIC*-оценкам в 3D-графике (рис. 1, справа). Единственное отличие в том, что локус 11 отсутствовал в Н-графе. Его индексы сходства с другими локусами графа были ниже 70 % (самый высокий с локусом 8 – 64,3 %).

Анализ соответствий (Correspondence Analysis, CA). Оценки *PIC* локусов варьировали от 0,222 до 0,825 (табл. 1). В пределах каждой выборки локусам присвоили одну из четырёх категорий: 1 – полиморфизм низкий ($PIC < 40\%$); 2 – средний ($PIC = 40\text{--}60\%$); 3 – высокий/значительный ($PIC = 61\text{--}80\%$); 4 – очень высокий ($PIC > 80\%$). Распределение локусов по категориям дано в таблице 4.

Таблица 4 – Число локусов по *PIC*-категориям /
Table 4 – The number of loci by *PIC*-categories

Порода / Breed	<i>PIC</i> -категория / <i>PIC</i> -category				Σ
	1	2	3	4	
JER	4	3	4	0	11
RED	0	4	6	1	11
HOL	0	3	6	2	11
Σ	4	10	16	3	33

Примечание / Note: 1 – $PIC < 40\%$; 2 – $PIC = 40\text{--}60\%$;
3 – $PIC = 61\text{--}80\%$; 4 – $PIC > 80\%$.



В JER-выборке отсутствовали локусы категории 4, в RED- и HOL-выборках – категории 1. Самая многочисленная во всех выборках категория 3. В относительных частотах (процентах) распределение представлено мозаичным графиком (рис. 3).

На графике высота каждой строки пропорциональна её общей частоте. Ширина полос в строке пропорциональна частоте каждой ячейки в этой строке. Площадь прямоугольника («бруска») пропорциональна частоте в конкретной ячейке таблицы 4.

Если локусы 3 и 4 категорий считать высокоинформативными, то в JER-выборке их было 36,4 %, в RED- и HOL – 63,6 и 72,7 % (при равном проценте локусов третьей категории – 54,5 %).

По построению и содержанию таблица 4 относится к таблицам сопряжённости, которые являются основой для CA¹⁴. Алгоритмы CA позволяют: (а) исследовать структуру и взаимосвязь группирующих объектов (строк) и переменных (столбцов); (б) представить содержимое таблицы в виде χ^2 -дистанции между отдельными объектами или/и переменными; (с) сформировать карту соответствия (correspondence map) в пространстве низкой размерности, на которой расстояние между двумя объектами (переменными) является мерой их сходства¹⁵.

Рис. 3. Мозаичный график распределения STR-локусов по категориям *PIC* /
Fig. 3. Mosaic plot of STR loci distribution by *PIC*-categories

В СА численности каждой ячейки таблицы сопряженности (n_{ij} в табл. 4) стандартизуются ($m_{ij} = n_{ij} / \Sigma n_{ij}$), чтобы сумма по всем ячейкам была равна 1. Полученная нормированная таблица показывает, как распределена единичная

масса по ячейкам. В терминологии СА суммы относительных частот (m_{ij}) по строкам и столбцам называют массой строки и столбца соответственно (m_i и m_j).

¹⁴Развитие алгебры СА началось с 1930-х годов [26], но до 1970-х годов метод был малоизвестным. По этой причине М. О. Хилл (M. O. Hill) [27] назвал его «забытым многомерным методом».

¹⁵В СА «соответствие» является мерой того, насколько сильно частоты между строками и столбцами таблицы данных отклоняются от нулевой модели – «отсутствие ассоциации».

Если строки и столбцы таблицы независимы друг от друга, то её элементы могут быть воспроизведены по профилям строк и столбцов (ожидаемая частота: $e_{ij} = m_i \cdot m_j / \sum n_{ij}$). При гипотезе о полной независимости значений по строкам и столбцам любое отклонение от ожидаемых величин ($r_{ij} = n_{ij} - e_{ij}$) будет давать вклад в совокупную [непараметрическую] статистику χ^2 Пирсона (фиксирует степень расхождения реальных частот от ожидаемых). Рассчитанная по таблице 4 статистика χ^2 представлена в таблице 5 (дистанция хи-квадрат для клетки: $d_{ij} = r_{ij}^2 / e_{ij}$).

Таблица 5 – χ^2 -дистанции /

Table 5 – Chi-Square Distances

Порода / Breed	PIC-категория / PIC-category				Σ
	1	2	3	4	
JER	5,33	0,03	0,33	1,00	6,70
RED	1,33	0,13	0,08	0,00	1,55
HOL	1,33	0,03	0,08	1,00	2,45
Σ	8,00	0,20	0,50	2,00	10,70

Комбинированный вклад всех ячеек в строке (столбце), Σ , рассматривается как *дистанция* между наблюдаемым профилем строки (столбца) и *центроидом* всех профилей строки (столбца), где центроид – это профиль, ожидаемый в предположении независимости. Совокупная χ^2 ($\Sigma d_{ij} = 10,70$) характеризует уровень соответствия значений строк и столбцов друг другу, или степень отклонения строк и столбцов от нулевой гипотезы.

При делении значений d_{ij} на Σd_{ij} полученные величины называют *относительной инерцией* (стандартизированное хи-квадрат; табл. 6). Суммирование по строке (столбцу) дает относительную инерцию строки (столбца). Последние можно представить точками-строками (точками-столбцами) в пространстве меньшей размерности, которое удовлетворительно воспроизводит сходство (и различие) между относительными частотами для строк (столбцов) таблицы сопряжённости.

Таблица 6 – Относительная инерция /

Table 6 – Relative Inertia

Порода / Breed	PIC-категория / PIC-category				Σ
	1	2	3	4	
JER	0,498	0,003	0,031	0,093	0,626
RED	0,125	0,012	0,008	0,000	0,145
HOL	0,125	0,003	0,008	0,093	0,229
Σ	0,748	0,019	0,047	0,187	1,000

Инерция характеризует *вариативность* (изменчивость) элементов таблицы сопряжённости. Общая инерция алгоритмами СА разделяется на последовательные компоненты – *размерности* (dimensions), которые не коррелируют друг с другом (также употребляется термин «axis» – ось [координат]). При этом максимизируется *соответствие* между оценками строк и оценками столбцов. То есть размерности (оси) выбираются так, чтобы расстояния между строками или столбцами были максимальными и увеличение числа осей приводило бы к уменьшению доли совокупной величины χ^2 (и, следовательно, инерционности), приходящейся на каждую дополнительную ось.

Относительная инерция размерности определяет вклад каждого объекта и каждой переменной в инерцию, приходящуюся на рассматриваемую размерность. Обычно первая размерность «захватывает» максимум инерции, вторая – следующую большую величину и т. д. Достаточно 2-3 главных (principal) размерностей, чтобы представить содержимое таблицы сопряженности в виде карты соответствия. На этой карте каждый объект или/и переменная изображаются точкой(-ами), позиции которых показывают, насколько сильно они «соответствуют» друг другу. Расстояние между двумя объектами (переменными) является мерой их сходства/различия.

Ординацию объектов (переменных) интерпретируют как аппроксимацию хи-квадрат дистанций между объектами (переменными). Чем больше инерция для конкретной размерности, тем выше вариативность среди профилей, которые представляют эту размерность. Таким образом, СА – это метод декомпозиции статистики χ^2 Пирсона с целью определения пространства наименьшей размерности, позволяющего представить отклонения от ожидаемых величин.

Для анализируемых данных 2D-карты соответствия показаны на рисунке 4 (А, В, С – по главным (principal) координатам, D – по стандартизированным (standard)).

На первую размерность (Dim1) пришлось 94,5 % извлеченной инерсии, на вторую (Dim2) – 4,5 %. Пересечение двух осей – это центр тяжести [инерции] наблюдаемых точек. Удаленность точек (выборок, категорий PIC) от центра тяжести указывает на их вклады в размерности (чем дальше, тем больше вклад).

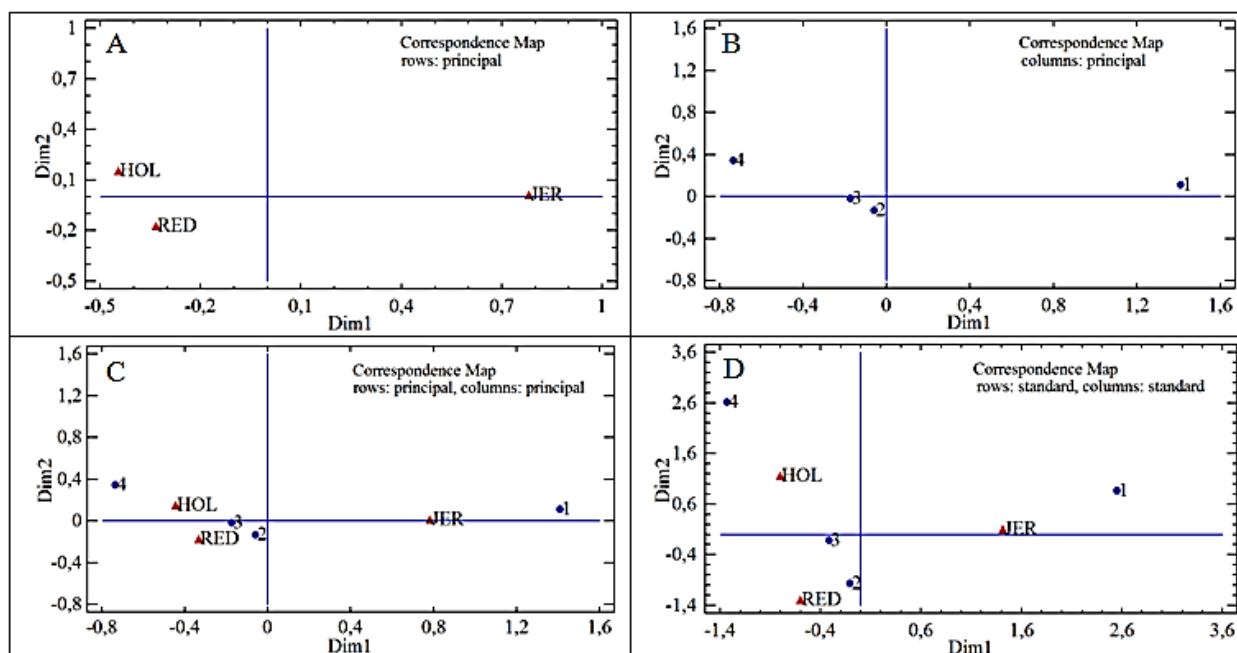


Рис. 4. 2D-карты соответствия: А – для выборок (▲); В – для категорий (●); С – биplot; D – биplot по стандартизованным координатам /

Fig. 4. 2D-correspondence maps: A – for samples (▲); B – for categories (●); C – biplot; D – biplot according to standardized coordinates

Интерпретация заключается в поиске группировок и контрастов в конфигурации проецируемых выборок и конфигурации проецируемых категорий. Расстояние/дистанция между однотипными точками (либо выборки, либо категории) характеризует их тесноту связи. Соотношение (корреляцию) между точкой (выборка, категория) и размерностью определяет угол с вершиной в центре тяжести. При остром угле – корреляция положительная, прямой угол говорит об отсутствии связи, тупой – об отрицательной корреляции.

Рисунок 4-А представляет карту соответствия породных выборок. Первая размерность отделила JER-выборку от двух других. Вторая размерность разнесла RED- и HOL-выборки: первая попала в III квадрант, вторая – во II квадрант. Характерна близость (схожесть по *PIC* профилям) RED- и HOL-выборок и их контраст с JER-выборкой. Точка «JER» находилась на значительном расстоянии от центра тяжести и, практически, лежала на горизонтальной оси. Поэтому вклад JER-выборки в инерцию Dim1 составил 66,2 %, в Dim2 – 0,4 %. Выборки RED и HOL имели более высокие вклады в Dim2 – 54,5 и 45,1 %, в Dim1 – 12,2 и 21,6 % соответственно.

Чем меньше угол между точкой-выборкой и осью, определяющей размерность, тем теснее связь между ними. Корреляция JER-выборки с Dim1 была равна +1 и нулю

с Dim2. Корреляции выборок RED и HOL с Dim1 были 0,79 и 0,89, с Dim2 – 0,21 и 0,11 соответственно. Между RED- и HOL-выборками можно полагать наличие положительной связи (острый угол с вершиной в центре тяжести) и отрицательной корреляции каждой из них с JER-выборкой (тупые углы).

Рисунок 4-В представляет 2D-карту соответствия категорий. Категории выстроились последовательно вдоль Dim1, которая отделила категорию 1 от остальных. Имели место близость 2 и 3 категорий (что указывает на сходство профилей), их контраст с категорией 4 и более значительный – с категорией 1. Категории 1 и 4 максимально не соответствовали из-за неоднородности данных в таблице сопряженности.

По горизонтальной оси точка 1 находится на значительном расстоянии от центра тяжести, но очень близка по вертикальной оси. Поэтому вклад категории 1 в инерцию Dim1 составил 78,6 %, в Dim2 – 9,2 %. На втором месте по величине вклада категория 4 – 16,2 %. Однако в Dim2 её вклад был наибольший – 62,3 %. Вклад категории 2 в Dim1 и Dim2 был соответственно 0,4 и 27,9 %, категории 3 – 4,9 и 0,6 %.

Корреляции категорий 1, 3 и 4 с Dim1 были 0,993, 0,993, 0,817 и почти нулевыми с Dim2 – 0,007, 0,007 и 0,183 соответственно.

Корреляция категории 2 с Dim1 была 0,183, с Dim2 – 0,817.

Угол между категориями 3 и 4 (через центр) был острым, что свидетельствовало о тесной положительной связи, которая объяснялась схожестью их профилей. Связности категорий 1 и 2, 2 и 3 не наблюдалось (углы практически прямые). Категории 1 и 3, 1 и 4 формировали тупые углы, указывающие на их отрицательное соотношение.

На рисунке 4-С дана комбинированная 2D-карта соответствия, представляющая всю информацию таблицы сопряженности (дистанционный биplot). Дистанции между объектами и переменными в биplotе точно не определяются. Однако небольшое расстояние указывает, примерно, на *повышенную вероятность* того, что данная переменная присутствует в данном объекте. Положения породных выборок вдоль Dim1 отражают их различия по конкретным категориям (проекция на Dim1). Выборки RED и HOL имели относительно высокие доли категорий 2 и 3, в то время как для выборки JER характерна категория 1. При этом для RED-выборки наиболее вероятны 2 и 3 категории локусов, для HOL-выборки – 3 и 4, для JER-выборки – 1 и 2 категории. Можно полагать, что имело место некоторое соответствие между расположением выборок и категорий с точки зрения их соотношения в матрице данных.

На рисунке 4-D представлен корреляционный биplot по опции «Plot standard coordinates». Опция полезна при интерпретации дистанций между строками (столбцами), координаты столбцов (строк) при этом не рассматриваются. По сравнению с картой соответствия 4-С, имела место некоторая поляризация точек-выборок и точек-категорий по Dim2, но их локализация относительно друг друга не поменялась.

Приблизительную связь между выборкой и категорией можно определить по углу с вершиной в центре тяжести. Если угол острый, то выборка и категория положительно коррелируют друг с другом. В нашем случае это означает, что JER-выборке соответствовал больший процент локусов первой категории, RED-выборке – 2 и 3 категорий, HOL-выборке – локусы категорий 3 и 4. Если угол тупой, то корреляция отрицательная (категории 3 и 4

для JER-выборки, категория 1 для RED-и HOL-выборок). Если угол прямой, то взаимосвязи нет (категория 2 для JER-выборки).

СА является исследовательским (разведочным) методом. Разработан на базе методологии, рассматривающей построение моделей с точки зрения их соответствия данным, а не наоборот [26]. Поэтому имело место мнение, что не существует тестов, которые могли бы быть использованы для проверки статистической значимости результатов СА. Однако некоторые оценки можно получить из таблицы 5. Статистика χ^2 – это наиболее простой критерий проверки значимости ассоциаций между категориальными (номинальными) переменными, не связанный с распределением последних. Тест χ^2 проверяет нуль-гипотезу (H_0): является ли классификация строк и столбцов независимой, т. е. принадлежность животных к породе не влияет на распределение STR-локусов по категориям. Рассчитанная величина χ^2 составила 10,7, а табличное значение критерия при числе степеней свободы $DF = (3 - 1)(4 - 1) = 6$ и уровне значимости $\alpha = 0,05$ было $\chi^2_{0,05; 6} = 12,59^{16}$ (больше, чем $\chi^2 = 10,7$). Также статистика *Likelihood ratio* составила 12,28 при $p_{value} = 0,056$ (больше, чем $\alpha = 0,05$). Оба теста указывали на то, что данные не противоречили модели независимости. Вместе с тем, если принять $\alpha = 0,10$, то табличное значение $\chi^2_{0,1; 6} = 10,64 < \chi^2 = 10,7$; при уровне доверия (доверности) 90 % гипотеза H_0 может быть отвергнута.

Коэффициент контингенции Чупрова, $T = \sqrt{\chi^2 / N \sqrt{DF}} = \sqrt{10,7 / 33 \sqrt{6}} = 0,36$, указывал на слабую степень зависимости между породными выборками и категориями STR-локусов. Коэффициент сопряженности Пирсона, рекомендованный в [28] для таблиц размером более чем 2×2 , составил $\phi = \sqrt{\chi^2 / N} = \sqrt{10,7 / 33} = 0,57$, что характеризовало связь как «относительно средняя». R-коэффициент Пирсона составил 0,417 ($p_{value} = 0,0158$), а τ_b -коэффициента Кендалла – 0,374 ($p_{value} = 0,0167$). Полученные оценки позволяют предположить наличие слабой, но статистически значимой связи между выборками и категориями PIC (влияния породы на распределение локусов по категориям).

¹⁶Находят по таблице критических значений χ^2 -распределения Пирсона в книгах по статистике, например: Гланц С. Медико-биологическая статистика. Электронная книга. М.: Практика. 1998. 459 с. Если $\chi^2_{0,05, DF} > \chi^2$, то H_0 принимается ($p_{value} > \alpha = 0,05$): строки и столбцы независимы.

Вероятности идентичности и исключения. Для некоторых исследований важна количественная оценка способности молекулярных маркеров распознавать разные индивиды. Эффективность идентификации оценивается вероятностью совпадения на основе данных о частотах аллелей и отношения генотипов по Харди-Вайнбергу. Наиболее часто используют *PI-статистику* (Probability Identity), которую интерпретируют как вероятность получения идентичных генотипов при определённом распределении частот аллелей¹⁷. Это может быть либо вероятность получения второго генотипа, идентичного тому, который уже есть, либо наличия в стаде животных с идентичными генотипами, либо получения двух идентичных генотипов в данной популяции. *PI-статистику* используют для сравнения потенциальной мощности различных локусов или разных наборов локусов, или для сравнения одного и того же набора локусов разных популяций, или для оценки необходимого числа локусов, обеспечивающих определённую надёж-

ность генетических тестов, или для идентификации индивидов по их гипервариабельным генетическим маркерам.

При допущении независимости аллелей и пропорций генотипов по Харди-Вайнбергу *PI* для каждого локуса (PI_L) рассчитывают по формуле: $PI_L = \sum p_i^4 + \sum (2p_i p_j)^2$, где p_i и p_j – частоты i -ой и j -ой аллели [29]¹⁸. Статистика PI_L представляет собой теоретическую вероятность идентичных генотипов локуса. Умножение PI_L на размер выборки даёт ожидаемое число животных с идентичными генотипами (Expected No. Individuals with the Same Genotype, ENISG).

Вероятность для возрастающей комбинации локусов (PI_C – мультилокусная вероятность) есть $PI_C = \prod PI_L$, где $\prod$ – оператор произведения (предполагает независимость аллелей разных локусов).

Перед расчётом PI_L , PI_C и *ENISG* (по PI_L) локусы были ранжированы по убыванию *PIC*. Ниже даны оценки по выборкам:

JER n = 10	<i>Eth225</i>	<i>Tgla227</i>	<i>Tgla122</i>	<i>Tgla53</i>	<i>Tgla126</i>	<i>inra23</i>	<i>Eth10</i>	<i>Bm1824</i>	<i>Eth3</i>	<i>Sps115</i>	<i>Bm2113</i>
PI_L	0,1200	0,1348	0,1463	0,1690	0,2131	0,2293	0,2310	0,4609	0,4609	0,5583	0,5875
ENISG	1,2	1,3	1,5	1,7	2,1	2,3	2,3	4,6	4,6	5,6	5,9
PI_C	0,1200	0,0162	0,0024	0,0004	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
RED n = 29	<i>Tgla227</i>	<i>Eth10</i>	<i>Eth225</i>	<i>Bm2113</i>	<i>Tgla53</i>	<i>Eth3</i>	<i>Tgla122</i>	<i>Tgla126</i>	<i>inra23</i>	<i>Bm1824</i>	<i>Sps115</i>
PI_L	0,0531	0,0698	0,0734	0,0862	0,0966	0,1289	0,1577	0,1757	0,2141	0,2398	0,3025
ENISG	1,5	2,0	2,1	2,5	2,8	3,7	4,6	5,1	6,2	7,0	8,8
PI_C	0,0531	0,0037	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
HOL n = 45	<i>Tgla53</i>	<i>Tgla227</i>	<i>Tgla122</i>	<i>Eth10</i>	<i>Bm2113</i>	<i>inra23</i>	<i>Bm1824</i>	<i>Eth225</i>	<i>Sps115</i>	<i>Eth3</i>	<i>Tgla126</i>
PI_L	0,0428	0,0519	0,0577	0,0685	0,0868	0,1076	0,1497	0,1559	0,1930	0,2001	0,2862
ENISG	1,9	2,3	2,6	3,1	3,9	4,8	6,7	7,0	8,7	9,0	12,9
PI_C	0,0428	0,0022	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Вероятность совпадения генотипов возрастала со снижением *PIC* локусов. В JER-выборке значения PI_L были в диапазоне [0,12; 0,5875] со средней **0,301**, в RED-выборке – [0,0531; 0,3025] со средней **0,145**, в HOL-выборке – [0,0428; 0,2862] со средней **0,127**. Животных с идентичными генотипами [«усреднённого» локуса] можно ожидать 30,0, 14,5 и 12,7 %; в абсолютных показателях,

с учётом численности выборок, – 3, 4 и 6 голов соответственно. Из трёх локусов с *min* значениями PI_L (максимальный *PIC*) локус *Tgla227* встречался во всех выборках, локус *Eth225* – в JER- и RED-выборках, локус *Tgla122* – в JER- и HOL-выборках, локус *Eth10* – только в RED-выборке, локус *Tgla53* – только в HOL-выборке.

¹⁷Статистика *PI* показывает среднюю вероятность того, что два неродственных индивида, взятых из одной и той же популяции со случайным спариванием, случайно окажутся носителями одного и того же мультилокусного генотипа (также называют *Population Match Probability* – популяционной вероятностью совпадения).

¹⁸В GenAlEx: $PI_L = 2(\sum p_i^2)^2 - \sum p_i^4$.

PI_C -статистика (по комплексу локусов) важна при планировании исследований, требующих точной идентификации индивидов. В частности, при определении числа высокоинформативных локусов, которые необходимы для достижения разумно низкого значения PI_C (например, 0,01–0,0001). Величину $(1 - PI_C)$ называют *мощностью исключения* (Exclusion Power), или *дискриминантной мощностью* (Power of Discrimination)¹⁹. Так, при идентификации животных для достижения мощности исключения на уровне $(1 - PI_C) = 0,9999$

(99,99 %) в JER-выборке достаточна комбинация из пяти, в RED-выборке – из четырёх, в HOL-выборке – из трёх высокоинформативных локусов.

Вероятность исключения (Probability Exclusion, PE). Обычно PE -статистику определяют, как вероятность исключения особи, как родителя, на основании несоответствия её генотипа с генотипом потомка хотя бы в одном из исследуемых локусов. Для отдельного локуса в программе GenAlEx имеются три статистики [30]:

$PE1$ – вероятность исключения, когда генотип одного родителя известен (the probability of exclusion when the other parent is known)

$$PE1 = 1 - 2 \sum p_i^2 + \sum p_i^3 + 2 \sum p_i^4 - 3 \sum p_i^5 - 2(\sum p_i^2)^2 + 3 \sum p_i^2 \sum p_i^3;$$

$PE2$ – вероятность исключения, когда генотип другого родителя известен, а другого – нет (the probability of exclusion when one parent is known, but the other genotype is unavailable)

$$PE2 = 1 - 4 \sum p_i^2 + 2(\sum p_i^2)^2 + 4 \sum p_i^3 - 3 \sum p_i^4;$$

$PE3$ – вероятность исключения пары предполагаемых родителей (the probability of excluding a putative parent pair)

$$PE3 = 1 + 4 \sum p_i^4 - 4 \sum p_i^5 - 3 \sum p_i^6 - 8(\sum p_i^2)^2 + 8(\sum p_i^2)(\sum p_i^3) + 2(\sum p_i^3)^2.$$

Вероятность исключения по комбинации локусов (мультилокусная) рассчитывается как $PE_C = 1 - \prod_{i=1}^{n_{loci}} (1 - PE_i)$, где PE_i – вероятность исключения, рассчитанная для каждого локуса ($PE1$, или $PE2$, или $PE3$). Эти статистики позволяют оценить среднюю способность маркерных систем выявлять несоответствия в записях родословных.

Результаты расчёта PE по отдельным локусам визуализированы на рисунке 5-А. Наиболее высокие оценки были по $PE3$ (исключение пары предполагаемых родителей); самые низкие – по $PE2$, когда известен генотип одного родителя. Оценки $PE1$ занимали промежуточное положение.

Обращает внимание близость оценок по RED- и HOL-выборкам и низкие оценки по JER-выборке. Так, усреднённые $PE1$ были для RED- и HOL-выборок 0,4767 и 0,5054 соответственно; в среднем 0,4910, или 49,1 %. По JER-выборке эта вероятность составила 30,4 %. Соответствующие оценки по $PE2$ были 32,1 и 17,1 %, по $PE3$ – 67,2 и 44,4 %.

Максимально возможные PE по отдельному локусу в JER-, RED- и HOL-выборках: $PE1$ – 49, 65, 69 %; $PE2$ – 31, 48, 53 %; $PE3$ – 67, 83 и 86 %. Величины явно недостаточные, особенно в JER-выборке, для надёжного тестирования отцовства/материнства.

На рисунке 5 (В, С, D) представлены кумулятивные кривые PE -статистик при возрастающей комбинации локусов. В частности, на рисунке 5-В даны $PE1$ -кумуляты по выборкам. При использовании только пяти самых информативных локусов кумулятивная $PE1$ по JER-выборке была 94,09 %, по RED-выборке – 98,89 %, по HOL-выборке – 99,34 %. Дополнительные шесть локусов повысили вероятности до 98,5 (+4,41), 99,94 (+1,05) и 99,97 (+0,63) % соответственно. Для достижения совокупной $PE1$ на уровне 99,9 %²⁰ в HOL-выборке было достаточно восемь, в RED-выборке – 10 локусов, в JER-выборке необходимо более 11 локусов.

¹⁹Мощность исключения – вероятность того, что два индивида будут иметь разные генотипы.

²⁰Из 1000 потомков, соответствующих своим родителям, можно ожидать только одного ложного совпадения (случайная ошибка).

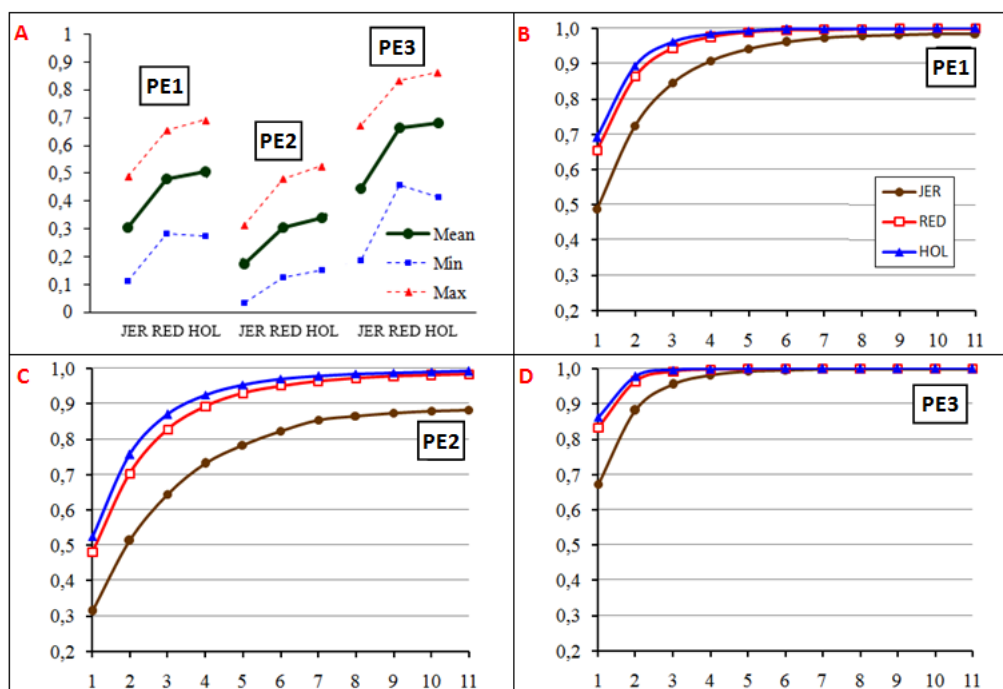


Рис. 5. Вероятности исключения: А – средние по 11 локусам с min и max значениями; В, С, D – кумулятивные для возрастающих комбинаций локусов (PE1 – когда известен генотип второго родителя; PE2 – когда генотип одного из родителей неизвестен; PE3 – когда исключаются оба предполагаемых родителя. В каждой выборке локусы были упорядочены по убыванию PIC) /

Fig. 5. Exclusion probabilities: A – the average for 11 loci with min and max values; B, C, D – cumulative for increasing combinations of loci (PE1 – when the genotype of the second parent is known; PE2 – when the genotype of one parent is unknown; PE3 – when both prospective parents are excluded. In each sample, the loci were ordered in descending order PIC)

Относительно мультилокусной PE2 (рис. 5-С; представляется более важной), то при использовании пяти локусов она была: в JER-выборке 78,19, в RED-выборке 93,08, в HOL-выборке 95,38 %. Комбинация из 11 локусов увеличила оценки PE2 до 88,26 (+10,07), 98,35 (+5,27) и 99,14 % (+3,76) соответственно. Прибавка была большей, чем в случае с PE1, но ни в одной выборке совокупная PE2 не достигла 99,9 %²¹. Ниже даны оценки кумулятивной PE2 для большего числа локусов, рассчитанные при допущении равного полиморфизма локусов (PIC из JER-выборки):

n loci →	11	15	20
PIC = 0,533	90,0	95,7	98,5
PIC = 0,611	95,4	98,5	99,6
PIC = 0,658	97,6	99,4	99,9
PIC = 0,680	98,3	99,6	99,9

Исходя из полученных результатов, для достижения в JER-выборке совокупной PE2 на уровне 99,9 % необходимо минимум

20 локусов, причём каждый из них должен иметь PIC не ниже 66 %.

При исключении обоих предполагаемых родителей, для достижения совокупной PE3 на уровне 99,9 %, в HOL-выборке было достаточно четырёх самых информативных локусов, в RED-выборке – пяти, а в JER-выборке – все 11 локусов (рис. 5-D). В случаях вольной случки или осеменения смешанной спермой, или подмены приплода PE3-тест является мощным средством идентификации кандидата в потомки.

Индивидуальная гетерозиготность.

У каждого производителя генотипы по 11 локусам были как в гетерозиготном, так и гомозиготном состоянии. Отношение числа гетерозиготных генотипов к их общему числу характеризовало индивидуальную гетерозиготность производителя (Individual Heterozygosity, H_{Ind}) [31]. Оценки H_{Ind} в JER-выборке были в диапазоне [0,273; 0,727], в RED-выборке – [0,455; 0,909] и в HOL-выборке – [0,455; 1]; средние 0,564, 0,734 и 0,814 соответственно.

²¹У речных буйолов (*Bubalus bubalis*) для достижения PE2 = 99,9 % потребовалось 11 локусов [32]. Для пород свиней крупная белая, дюрок и ландрас совокупные PE2 по 10 STR локусам были 96,4, 93 и 90,5 % соответственно [33].

По *PIC* локусов каждая выборка была разделена на две группы: «*PIC+*» – субвыборка животных с пятью высокоинформативными локусами, «*PIC-*» – с пятью низкоинформативными локусами²². В каждой субвыборке были рассчитаны новые H_{Ind} -оценки производителей, после чего данные были объединены и обрабо-

таны ANOVA по статистической модели фиксированного типа: $H_{Ind} = \mu + Breed + grPIC + e$ (табл. 7).

Влияние факторов «*Breed*» и «*grPIC*» было статистически значимым. Их суммарный вклад в общую изменчивость H_{Ind} составил 28,2 %. На фактор «*Breed*» приходилось 12,3 %²³, что указывало на низкую ($\sqrt{0,123} = 0,35$) ассоциацию с H_{Ind} .

Таблица 7 – Результаты ANOVA по H_{Ind} /

Table 7 – Results ANOVA for H_{Ind}

Source	SS	η^2 , %	DF	MS	F-test	p-value
Breed	1,292	12,3	2	0,646	14,01	<0,0001
grPIC	1,680	15,9	1	1,680	36,42	<0,0001
Residual (e)	7,564	71,8	164	0,046	-	-
Total (corrected)	10,536	100,0	167	-	-	-

Среднее H_{Ind} в JER-выборке было статистически значимо ниже, чем в RED- и HOL-выборках – их 95 % CI не перекрывались (рис. 6, слева). Различие между средними RED- и HOL-выборок было статистически незначимым.

В пределах субвыборок («*PIC+*» или «*PIC-*») 95 % CI H_{Ind} -средних не перекрывались только у производителей JER- и HOL-пород (рис. 6, справа). В пределах пород различие H_{Ind} -средних между «*PIC+*» и «*PIC-*» было, примерно, равным (в «*PIC-*» на 25 %

ниже), но в JER-выборке статистически незначимым (95 % CI перекрывались).

Корреляция Пирсона между H_{Ind} -оценкам производителей, рассчитанных по субвыборкам «*PIC+*» и «*PIC-*», была «слабой» – $r = 0,266$ ($p_{value} = 0,0143$) с 95 % CI [0,055; 0,455]; *bootstrapping* дал 95 % CI [0,115, 0,437]²⁴. Исходя из 95 % CI, максимальная повторяемость H_{Ind} -оценок не превышала 20 %. Ранговые корреляции Спирмена и Кендалла были 0,316 и 0,267 соответственно.

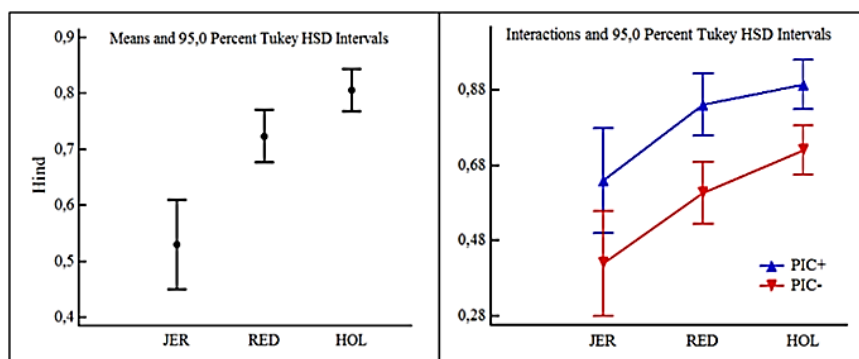


Рис. 6. LS-средние H_{Ind} с 95 % CI по выборкам (слева) и группам по *PIC* (справа) /
Fig. 6. Least Squares Means H_{Ind} and 95 % CI by samples (left) and by *PIC*-groups (right)

Влияние уровня *PIC* на оценку генетической дистанции. По RED + HOL выборке²⁵ были сформированы три субвыборки, представляющих одних и тех же производителей, но с разными наборами локусов. Субвыборка PIC_{max} включала четыре высокоинформативных локуса (*Tgla227*, *Tgla53*, *Bm2113*

и *Eth10*), субвыборка PIC_{min} – четыре низкоинформативных локуса (*Eth3*, *Tgla126*, *Sps115*, *Bm1824*) и субвыборка PIC_{mix} (смешанная) – остальные три локуса с [условно] высоким, средним и низким показателями *PIC* (*inra23*, *Tgla122* и *Eth225*).

²²Одни и те же производители, но с разными наборами локусов.

²³Аналогичный результат по однофакторной модели.

²⁴Howell D. C. Resampling Procedures. University of Vermont. Copyright 2000 (Package).

²⁵JER-выборка выпала из-за малого числа локусов с высокими показателями *PIC*.

По каждой субвыборке и исходной («Total» с 11-ю локусами) рассчитали индексы фиксации (G_{ST} , $G'_{ST(N)}$), индексы фиксации, модифицированные в индексы дифференциации ($G'_{ST(H)}$, G''_{ST}), и «подлинные» индексы дифференциации (D_{est} , D'). Оценки по Total-выборке приняли за «истинные» и использовали для расчёта MAPE (Mean Absolute Percentage Error)²⁶ – средней абсолютной процентной ошибки пары *подобных* статистик,

полученных по той или иной субвыборке. Результаты сведены в таблице 8 и визуализированы на рисунке 7.

Прежде всего следует отметить, что по Total-выборке «истинные» оценки индексов фиксации были в 3 и более раз ниже индексов дифференциации (включая модифицированные). Похожее отношение сохранялось и для оценок по трём субвыборкам.

Таблица 8 – Генетические дистанции RED-HOL, рассчитанные по субвыборкам с разным уровнем PIC локусов шестью методами /

Table 8 – Genetic distances between RED and HOL, calculated from subsamples with different levels of PIC loci using six methods

Мера / Measure → Вариант / Variant ↓	Fixation indices		Modified indices		Differentiation indices	
	G_{ST}	$G'_{ST(N)}$	$G'_{ST(H)}$	G''_{ST}	D_{est}	D'
Total (11 locus)	0,056	0,105	0,331	0,366	0,292	0,343
PIC_{max} (4 locus)	0,034	0,066	0,319	0,342	0,295	0,355
PIC _{min} (4 locus)	0,070	0,130	0,292	0,338	0,238	0,269
PIC _{mix} (3 locus)	0,070	0,131	0,430	0,468	0,388	0,421
MAPE (PIC _{max}), %	38,2		5,1		2,3	
MAPE (PIC _{min}), %	24,4		9,7		20,0	
MAPE (PIC _{mix}), %	24,9		28,9		27,8	

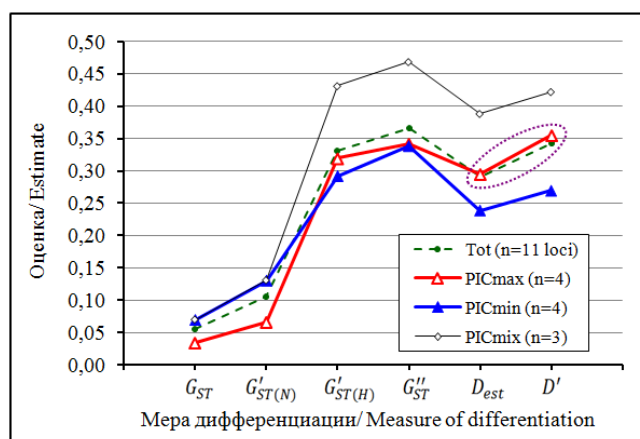


Рис. 7. Оценки генетической дистанции RED-HOL на выборках с разным уровнем PIC локусов /

Fig. 7. Estimates of the RED-HOL genetic distance in samples with different levels of PIC loci

Абсолютное расхождение оценок, рассчитанных по субвыборкам PIC_{max} и PIC_{min}, для индексов фиксации было в среднем 0,05, для модифицированных индексов – 0,016, для «подлинных» индексов дифференциации – 0,072. Оценки индексов фиксации по PIC_{max}-субвыборке были в 2 раза *ниже*, чем по PIC_{min}-субвыборке. В то время как оценки индексов дифференциации по PIC_{max}-субвыборке были *выше* таковых по PIC_{min}-субвыборке на 16,6 %. При этом оценки по $G'_{ST(H)}$

и G''_{ST} мерам были выше на 5,2 %, а по D_{est} и D' мерам – на 28 %.

Интересно, что индексы дифференциации, рассчитанные по PIC_{mix}-субвыборке (смешанной), были на 19–63 % выше, чем аналогичные индексы по другим субвыборкам. Как представляется, это могло быть следствием неоднородности локусов по оценкам (точнее, по числу [и частоте] аллелей). Однако неоднородность данных не влияла на индексы фиксации, так как оценки, рассчитанные по PIC_{mix} и PIC_{min} субвыборкам, совпадали.

²⁶MAPE = (100 × Σ|e_i|/θ*)/n, где e_i = θ_i - θ*; θ_i – i-ая оценка, θ* – «истинная» оценка.

Индексы фиксации, рассчитанные по трём субвыборкам, имели MAPE 24–38 %. Оценки индексов дифференциации по субвыборке PIC_{max} имели самые малые MAPE – в среднем 3,7 % (по субвыборке PIC_{min} – 10–20 %). При использовании субвыборки PIC_{mix} оценки индексов фиксации и дифференциации имели MAPE на уровне 25–29 %.

В целом, уровень PIC локусов влиял на оценки относительной генетической дифференциации выборок. Ближе к «истинным» были оценки, полученные по субвыборке с четырьмя высокоинформативными локусами мерами D_{est} , D' (MAPE = 2,3 %), и $G'_{ST(H)}$, G''_{ST} (MAPE = 5,1 %). В остальных вариантах MAPE были >10%, что можно интерпретировать как статистически значимое отклонение от «истинных» оценок.

Заключение. Породы анализируемых выборок производителей имели разную историю. Так, джерсейская порода разводилась на острове Джерси в изоляции 2,5 века (1763–2008 гг.). Обмен генетическим материалом между финской айрширской, красной датской и краснопёстрой шведской породами, продолжавшийся в течение многих лет, способствовал получению потомства различной кровности, смешиванию генофондов и, как следствие, объединению (2010 г.) в популяцию Viking Red – красную скандинавскую породу. Начавшаяся в 70-х годах XX столетия экспансия голштинской породы Северной Америки в европейские, а затем и в другие страны мира, сделали её трансграничной²⁷. Рассматриваемая HOL-выборка включала данные по голштинским производителям североамериканской, голландской и немецкой селекции. Особенности разведения (в JER-породе – инбридинг, в RED – трёхпородное скрещивание+селекция, в HOL – поглотительное скрещивание+селекция) сформировали генофонды пород, отличающиеся по показателям аллельного и генного разнообразия. Самыми низкими они были в JER-выборке, самыми высокими – в HOL-выборке. Так, по JER-выборке число эффективных аллелей на локус (S_{ne}) было 2,7, в RED-выборке – 4,4 и в HOL-выборке – 4,8; ожидаемая гетерозиготность (H_e) соответственно 0,533, 0,683 и 0,724, информационный контент полиморфизма (PIC) – 0,470, 0,650, 0,682, причём оценки были статистически значимо ниже H_e -оценок. Средние S_{ne} , H_e и PIC показатели

по JER-выборке статистически значимо отличались от таковых в RED и HOL выборках. Между последними статистически значимых различий не наблюдалось (по всей вероятности, вследствие кроссбридинга как при слиянии, так и при поглощении пород). В пределах породных выборок имела место сильная взаимосвязь PIC -оценок с оценками S_{ne} и H_e (0,940–0,998). В то же время, корреляции PIC -оценок локусов между породными выборками были статистически незначимые.

PIC -оценки в JER-выборке имели высокую изменчивость ($CV = 37$ %); выборки RED и HOL были более гомогенными ($CV = 17$ %). Локусов с $PIC > 0,6$ в JER-выборке, примерно в 2 раза меньше, чем в RED- и HOL-выборках (36,4, 63,6 и 72,7 %), что характеризовало последние как генетически высокополиморфные. Из трёх локусов с тах PIC локус $Tgla227$ встречался во всех выборках, локус $Eth225$ – в JER- и RED-выборках, локус $Tgla122$ – в JER- и HOL-выборках, локус $Eth10$ – только в выборке RED, локус $Tgla53$ – только в выборке HOL. Сетевой анализ с ограничением для индекса сходства локусов $S \geq 50$ % сформировал граф размером 26 межлокусных связей ($\bar{S} = 61,6$ %) из 55 возможных ($\bar{S} = 47,6$ %). Локусы 1, 2, 5, 7, 8 и 11 имели по 5–7 связей («влиятельные»). Повышенные индексы сходства были в цепях 3-8-5 и 6-10 (≈ 80 %). При условии $S \geq 70$ % компоновка сети включала четыре нулевых графа (локусы 4, 11, 7 и 9) и два цепных 6-10-1 и 3-8-5-2. Первый граф объединил низкополиморфные локусы с $\overline{PIC} = 0,472$, второй – высокополиморфные с $\overline{PIC} = 0,686$. Сходную группировку наблюдали в трёхмерном (JER×RED×HOL) графике рассеяния локусов по PIC .

Близость выборок RED и HOL и их контраст с JER-выборкой наглядно демонстрировала 2D-карта соответствия (CA-анализ). Имела место некоторая сопряжённость между ординациями выборок и категорий (1 категория – $PIC < 40$ %; 2 – $PIC = 40–60$ %; 3 – $PIC = 61–80$ %; 4 – $PIC > 80$ %). Гипотеза о независимости принималась при $\alpha = 0,05$ и отклонялась при $\alpha = 0,1$ (тесты χ^2 и *Likelihood ratio*). Статистически значимые оценки корреляций Пирсона и Кендалла ($\approx 0,4$) позволили предположить наличие слабой ассоциации между породой производителя и PIC -категорией.

²⁷Порода, которая разводится в нескольких странах и регионах (международная) или в нескольких странах, но внутри одного региона (региональная).

Оценки индивидуальной гетерозиготности производителей (H_{Ind}) варьировали от 0,273 до 0,727 в JER-выборке, от 0,455 до 0,909 в RED-выборке и от 0,455 до 1 в HOL-выборке при средних 0,564, 0,734 и 0,814 соответственно. Влияние породной принадлежности и групп с высоким и низким полиморфизмом локусов на изменчивость H_{Ind} было статистически значимым и, примерно, равным (12,3 и 15,9 %). Средние H_{Ind} по породам в субвыборке «PIC-» были на ~25 % ниже, но различие в JER-породе статистически незначимо. Взаимосвязь H_{Ind} -оценок одних и тех же производителей, рассчитанных по высокоинформативным и низкоинформативным локусам, была «слабой»; максимальная повторяемость H_{Ind} -оценок не превышала 20 %.

Уровень PIC локусов влиял на оценку относительной генетической дифференциации RED- и HOL-выборок. Полученные результаты в целом соответствовали теоретическим разработкам, демонстрирующим, что при анализе мультиаллельных маркеров повышенный уровень гетерозиготности в пределах популяций занижает оценки межпопуляционного разнообразия, измеряемого индексами фиксации [18, 34; см. также 20, 21]. В то же время заниженные оценки генетической дифференциации «правильными» методами по PIC_{min} -выборке (низкоинформативные локусы) сигнализировали о том, что использование ДНК-маркеров с малой аллельностью (например, SNP) может привести к ошибочным заключениям. Аналогично и при использовании локусов, неоднородных по PIC. По всей вероятности, использование в популяционно-генетических исследованиях высоко-

информативных ДНК-маркеров совместно с D_{est} и D' мерами должно способствовать получению объективных оценок относительной генетической дифференциации популяций.

Как представляется, при идентификации происхождения животных на базе STR-локусов могут быть полезны разведочные исследования по (1) оцениванию PIC локусов тестируемой выборки и (2) анализу необходимого числа локусов, которые обеспечили бы требуемую вероятность исключения предполагаемого родителя(-ей). Во всяком случае, результаты, полученные по данным JER- и RED-выборкам, свидетельствовали о том, что для тестирования отцовства/материнства с совокупной $PE2 = 99,9$ % [известен генотип одного родителя] имеющихся 11 локусов было недостаточно. В общем, вероятность исключения 99,9 % может быть обеспечена с меньшим числом локусов, если для PIC устанавливать ограничение, например, не ниже 60 %.

Итак, выявляя и выбирая высокоинформативные STR-локусы [и ДНК-маркеры] можно повысить надежность тестов и генетических оценок, снизить вероятность ошибок при индивидуальной, породной и видовой идентификации, оценке родства, а также исключить риски при проверке происхождения животных. Такой подход будет способствовать сокращению числа необходимых локусов и, соответственно, затрат при проведении масштабных популяционно-генетических исследований. Существует мнение, что инвестиции в поиск высокоинформативных локусов могут быть экономически выгодными в долгосрочной перспективе [35].

References

1. Nei M. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences. 1973;70(12):3321–3323. DOI: <https://doi.org/10.1073/pnas.70.12.3321>
2. Botstein D., White R. L., Skolnick M., Davis R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics. 1980;32(3):314–331. URL: <https://pubmed.ncbi.nlm.nih.gov/6247908/>
3. Speer M. J. Genetic linkage: concepts and methods. In: Albers MJ (ed) Genetics of cerebrovascular disease. Blackwell, Oxford, 1999. pp. 25–26.
4. Guo X., Elston R. C. Linkage informative content of polymorphic genetic markers. Human Heredity. 1999;49(2):112–118. DOI: <https://doi.org/10.1159/000022855>
5. Yasuda N. HLA Polymorphism Information Content (PIC). Japanese journal of human genetics. 1988;33(3):385–387. DOI: <https://doi.org/10.1007/BF02032870>
6. Chakravarti A. Information content of the Centred Etude du Polymorphisme Humain (CEPH) family structures for linkage studies. Human Genetics. 1991;87:721–724. URL: <https://link.springer.com/article/10.1007/BF00201732>
7. Hildebrand C. E., Torney D. C., Wagner R. P. Informativeness of polymorphic DNA markers. In: Mapping the Genome/Informativeness of Polymorphic DNA Markers. Los Alamos Science. 1992;20:100–102. URL: <https://sgp.fas.org/othergov/doe/lanl/pubs/00326695.pdf>

8. Da Y., VanRaden P. M., Ron M., Beever J. E., Paszek A. A., Song J. et al. Standardization and conversion of marker polymorphism measures. *Animal Biotechnology*. 1999;10(1-2):25–35.
DOI: <https://doi.org/10.1080/10495399909525919>
9. Веллер Д. И. Геномная селекция животных. СПб.: Проспект Науки, 2018. 208 с.
Weller J. I. Genomic breeding of animals. Saint-Petersburg: *Prospekt Nauki*, 2018. 208 p.
10. Höschele I. Mapping quantitative trait loci in outbred pedigrees. In: *Handbook of Statistical Genetics*, Third Edition. Edited by D. J. Balding, M. Bishop and C. Cannings. John Wiley & Sons, Ltd. 2007. pp. 623–677.
11. Gilbert J. E., Lewis R. V., Wilkinson M. J., Caligari P. D. S. Developing an appropriate strategy to assess genetic variability in plant germplasm collections. *Theoretical and Applied Genetics*. 1999;98(6-7):1125–1131.
DOI: <https://doi.org/10.1007/s001220051176>
12. Столповский Ю. А., Бабаян О. В., Каштанов С. Н., Пискунов А. К., Сёмина М. Т., Холодова М. В. и др. Генетическая оценка пород северного оленя (*Rangifer tarandus*) и их дикого предка с помощью новой панели STR-маркеров. *Генетика*. 2020;56(12):1410–1426. DOI: <https://doi.org/10.31857/S0016675820120139> EDN: YCPKAN
- Stolpovskiy Yu. A., Babayan O. V., Kashtanov S. N., Piskunov A. K., Se-mina M. T., Kholodova M. V. et al. Genetic evaluation of the breeds of reindeer (*Rangifer tarandus*) and their wild ancestor using a new panel of STR markers. *Genetika* = Russian Journal of Genetics. 2020;56(12):1410–1426. (In Russ.).
DOI: <https://doi.org/10.31857/S0016675820120139>
13. 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>
14. 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>
15. 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>
16. Hartl D. E., Clark A. G. *Principles of population genetics*. 3rd Ed. Sinauer Associates. Sunderland, Mass. 1997. 543 p.
17. Meirmans P. G., Hedrick P. W. Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources*. 2011;11(1):5–18. DOI: <https://doi.org/10.1111/j.1755-0998.2010.02927.x>
18. Jost L. G_{ST} 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>
19. Smouse P. E., Banks S. C., Peakall R. Converting quadratic entropy to diversity: Both animals and alleles are diverse, but some are more diverse than others. *PLoS ONE*. 2017;12(10):e0185499.
DOI: <https://doi.org/10.1371/journal.pone.0185499>
20. 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: FYQNTTE
- 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>
21. Кузнецов В. М. Информационно-энтропийный подход к анализу генетического разнообразия популяций. *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
- 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>
22. 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>
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. 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.
URL: https://palaeo-electronica.org/2001_1/past/past.pdf
25. STATGRAPHICS® Centurion XVI User Manual. By StatPoint Technologies, Inc. 2010. 297p.
26. Greenacre M. J. *Theory and applications of correspondence analysis*. London: Academic Press, 1984. 364 p.
27. Hill M. O. Correspondence analysis: a neglected multivariate method. *Appl. Statist.* 1974;23(3):340–354.
DOI: <https://doi.org/10.2307/2347127>

28. Хромов-Борисов Н. Н. Корректный статистический анализ частот генотипов по генам семейств UCP и PPAR в группе жителей блокадного Ленинграда и контрольной группе. Успехи геронтологии. 2016;29(3):454–460. Режим доступа: <https://www.elibrary.ru/item.asp?id=26744152> EDN: WNGHOH

Khromov-Borisov N. N. correct statistical analysis of genotype frequencies for UCP and PPAR gene families in residents of besieged Leningrad and the control group. *Uspekhi gerontologii*. 2016;29(3):454–460. (In Russ.). URL: <https://www.elibrary.ru/item.asp?id=26744152>

29. Paetkau D., Strobeck C. Microsatellite analysis of genetic variation in black bear populations. *Molecular Ecology*. 1994;3(5):489–495. DOI: <https://doi.org/10.1111/j.1365-294x.1994.tb00127.x>

30. Jamieson A., Taylor S. C. Comparisons of three probability formulae for parentage exclusion. *Animal Genetics*. 1997;28(6):397–400. DOI: <https://doi.org/10.1111/j.1365-2052.1997.00186.x>

31. Waits L. P., Luikart G., Taberlet P. Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology*. 2001;10(1):249–256. DOI: <https://doi.org/10.1046/j.1365-294x.2001.01185.x>

32. Kathiravan P., Kataria R. S., Mishra B. P. Power of exclusion of 19 microsatellite markers for parentage testing in river buffalo (*Bubalus bubalis*). *Molecular Biology Reports*. 2012;39(8):8217–8223. DOI: <https://doi.org/10.1007/s11033-012-1669-x>

33. Харзинова В. Р., Костюнина О. В., Карпушкина Т. В., Зиновьева Н. А. Способ определения достоверности происхождения и чистопородности свиней: пат. №2744733 Российская Федерация. №2019142089: заявл. 18.12.2019; опублик. 15.03.2021. Бюл. № 8. 19 с.

Режим доступа: https://www1.fips.ru/registers-doc-view/fips_servlet

Kharzinova V. R., Kostyunina O. V., Karpushkina T. V., Zinoveva N. A. Method for determination of origin and genuineness of strain of pigs: Patent RF, no.2744733, 2021.

URL: https://www1.fips.ru/registers-doc-view/fips_servlet

34. Hedrick P. W. Perspective: Highly variable loci and their interpretation in evolution and conservation. *Evolution*. 1999;53(2):313–318. DOI: <https://doi.org/10.1111/j.1558-5646.1999.tb03767.x>

35. Lynch M., Ritland K. Estimation of pairwise relatedness with molecular markers. *Genetics*. 1999;152(4):1753–1766. DOI: <https://doi.org/10.1093/genetics/152.4.1753>

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

✉ Vasily 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