МЕТОДЫ

УДК 575.174:636.082 *DOI: 10.25687/1996-6733.prodanimbiol.2020.1.91-110*

МЕТОДЫ НЕЯ ДЛЯ АНАЛИЗА ГЕНЕТИЧЕСКИХ РАЗЛИЧИЙ МЕЖДУ ПОПУЛЯЦИЯМИ

Кузнецов В.М.

Федеральный аграрный научный центр Северо-Востока им. Н.В. Рудницкого, Киров, Российская Федерация

Ключевым вопросом при определении и измерении дифференциации любых популяций является количественная оценка неслучайного распределения генетической изменчивости. Исследования дивергенции видов и генетической дифференциации популяций требуют анализа и гетерозиготности (разнообразия) и генетических расстояний (дистанций), которые измеряют разные аспекты изменчивости. Знание того, как генетическая изменчивость распределяется между популяциями, имеет важные последствия не только для эволюционной биологии и экологии, но и для разведения и сохранения пород продуктивных животных. Имеются различные методы и компьютерные программы для анализа генетической изменчивости по маркерам ДНК (микросателлитам, однонуклеотидному полиморфизму), которые используются при исследовании популяций животных. Вместе с тем генетико-математические основы методов в российских публикациях отражены недостаточно. Их рассмотрение и являлось целью настоящей работы. В частности, представлены подходы Нея (Nei, 1974-1994) к оценке генетических различий между популяциями, базирующиеся на вероятности идентичности случайно извлечённых генов в пределах и между популяциями. В отличие от индекса фиксации Райта для диаллельного локуса, статистики Нея выражены в терминах внутри- и межпопуляционного генного разнообразия. Представлены формулы расчёта парных генетических дистанций и сводных оценок генной дифференциации популяций. На численных примерах иллюстрируются: предварительный χ^2-тест различия аллельных профилей популяций, расчёты несмещённых оценок минимальной (uD_{min}), стандартной (uD_N) и максимальной (uD_{max}) генетических дистанций, сводных оценок абсолютной (uD_{ST}) и относительной (uG_{ST}) генной дифференциации, их варианс и стандартных ошибок. Меры генного разнообразия Нея применимы к любым популяциям независимо от числа локусов, полиморфности аллелей в локусе, наличия эволюционных факторов (мутаций, миграции, дрейф генов и отбора). Оценки генной дифференциации и генетических дистанций Нея по молекулярно-генетическим маркерам могут служить ценной дополнительной информацией, позволяющей селекционерам в совокупности с традиционными и биометрическими методами принимать правильные решения по разведению, улучшению, кроссбридингу и сохранению пород продуктивных животных.

Ключевые слова: гетерозиготность, генетическое разнообразие, генетическая дистанция. коэффициент генной дифференциации

Проблемы биологии продуктивных животных, 2020, 1: 91-110

Введение

Сходство или различие природных популяций по типу, степени и характеру генетической изменчивости могут быть результатом комплекса факторов. Так, генетическое сходство может быть обусловлено тем, что популяции только начали дивергировать, или между ними существует поток генов (миграция), или имеет место незначительный дрейф генов из-за их большой численности, или отбор в одинаковой степени влияет на схожие локусы. Различия между популяциями могут быть вызваны долговременной изоляцией и отсутствием миграции, или случайным генетическим дрейфом, или дифференцированным отбором. В реальных условиях могут действовать несколько или даже множество факторов (Hedrick, 2003). В разведении продуктивных животных генетические различия между породами, линиями, стадами (хозяйствами) могут быть следствием разных целей и интенсивности селекции, численности поголовья и масштаба использования «лучшего мирового генофонда», систем воспроизводства (используется или нет искусственное осеменение, трансплантация

эмбрионов, сексированная сперма). Исследования дивергенции видов и генетической дифференциации популяций требуют анализа и гетерозиготности (разнообразия) и генетических расстояний (различий), которые измеряют различные аспекты изменчивости. Аллельное (нуклеотидное) разнообразие и гетерозиготность оценивают взвешенную изменчивость особей в популяциях, тогда как дистанция/сходство и сводные коэффициенты дифференциации измеряют попарные или групповые (сводные) различия между популяциями по маркерным генам или молекулярным последовательностям. Знание того, как генетическая изменчивость распределяется между популяциями, имеет важные последствия не только для эволюционной биологии и экологии, но и для разведения и сохранения пород сельскохозяйственных животных. В частности, надёжные оценки генетических дистанций и коэффициентов дифференциации имеют решающее значение для понимания генетических отношений между популяциями (группами) животных и представляют собой показатели, необходимые при разработки стратегий разведения пород и сохранения генофондных стад.

Согласно неодарвинистским воззрениям, новый вид возникает в результате дифференциации любой популяции, относящейся к нему, которая проявляется в постепенной дивергенции на уровне генофонда. В основе дивергенции лежит процесс постепенной замены одних аллелей определённых генов на другие. В процессе дивергенции двух популяций всё меньшее число генов имеет аллели, которые встречаются в обеих популяциях. Когда заканчивается процесс видообразования, для всех генов существуют аллели, характерные только для одной из популяций. При этом генетическое сходство становится равным нулю и новый вид по морфологическим признакам становится отличным от первоначального (https://ru.wikipedia.org/wiki/Неодарвинизм). По нейтральной теории молекулярной эволюции, большая часть мутационных замещений в ходе эволюции обусловлена не положительным дарвиновским отбором, а случайным закреплением нейтральных или почти нейтральных мутаций (Kimura, 1985). Внутривидовая молекулярная генетическая изменчивость, проявляющаяся в виде полиморфизма белков, селективно нейтральна или почти нейтральна. Этот полиморфизм поддерживается в популяциях любого вида благодаря равновесию между мутационным процессом и случайной элиминацией или фиксацией аллелей. Основными факторами молекулярной эволюции являются мутационный процесс и случайный дрейф генов.

Для изучения генетических процессов в подразделённых популяциях Райт предложил три индекса фиксации: F_{IS} – коэффициент инбридинга индивидов внутри субпопуляций, F_{IT} – коэффициент инбридинга индивидов в объединённой популяции и F_{ST} – коэффициент межсубпопуляционных генетических различий (Wright 1943, 1951; см. также Kuznetsov, 2014). Индексы базировались на однолокусной диаллельной модели популяции, что делало их применение проблематичным. В 1970-х годах Ней (Nei, 1971, 1972, 1973, 1977, 1978, 1987) модифицировал индексы фиксации Райта, предложив иной подход к исследованию подразделённых популяций. Ней (Nei, 1973) показал, что генное разнообразие во всей популяции может быть разложено на две компоненты: внутри- и межсубпопуляционное генное разнообразие, если генное разнообразие понимать как гетерозиготность по Харди-Вайнбергу. В теории Нея генное разнообразие определяется путём использования генных частот текущей генерации, поэтому нет необходимости в предположениях о родословных индивидов, отборе и миграции в прошлом. Генетические дистанции и сводные коэффициенты генной дифференциации по Нею не зависят от плоидности организмов (диплоидные или полиплоидные), репродуктивной системы (половое или бесполое размножением), количества и численности субпопуляций. Для методов Нея характерны «простая формулировка, легкость применения и ясность биологического смысла» (Kimura, 1985).

Меры генного разнообразия Нея наиболее часто используются в био-зоотехнических исследованиях изменчивости ДНК-маркеров, но их описание в российских публикациях освещено недостаточно. Цель данной работы – рассмотрение генетико-математической основы таких статистик Нея, как минимальная, стандартная и максимальная генетические дистанции, коэффициенты абсолютной и относительной генной дифференциации популяций. Для лучшего понимания методов представлялось целесообразным дать гипотетические примеры, иллюстрирующие процедуры расчёта этих статистик.

Проверка выборок по χ^2-критерию

С теоретической точки зрения, статистики Нея должны вычисляться по популяционным частотам аллелей всех локусов генома, но на практике такое обследование невозможно. Как правило, статистики Нея оценивают по случайно выбранному из популяций некоторому числу особей и анализу некоторого случайного числа локусов. Поэтому оценке статистик должны предшествовать два процесса *рандомизированного* отбора: особей (генов) из популяций и локусов из генома.

Перед анализом генетического разнообразия выборок рекомендуют определять статистическую значимость их различий по выборочным частотам аллелей каждого локуса (Weir, 1995). Для этого используют χ^2- критерий (Workman, Niswander, 1970). χ^2-критерий для простого локуса (число выборок s = 2; число локусов (генов) m = 1, число аллелей (аллельных состояний гена) r = 2; число степеней свободы df = (s-1)(r-1) =1):

$$\chi^2 = \frac{\sum_i^s (2n_i)\, p_i^2 - \bar{p} \sum_i^s (2n_i)\, p_i}{\bar{p}\,\bar{q}},$$

где p_i - относительная частота p-ой аллели в i-ой выборке и n_i - число особей в i-ой выборки; $N=\sum n_i$; $\bar{p} = \sum_i^s n_i p_i / N$; $\bar{q} = 1 - \bar{p}$.

Адекватно соотношение:

$$\chi^2 = 2N \frac{\sum_i^s w_i\, p_i^2 - \bar{p}^2}{\bar{p}\,(1-\bar{p})},$$

где $w_i = n_i / N$ – «вес» для i-ой выборки.

Выражение в числителе – это взвешенная варианса p_i:

$$\sum_i^s w_i\, p_i^2 - \bar{p}^2 \;= \sum_i^s w_i (p_i - \bar{p})^2 = \sigma_p^2.$$

Тогда:

$$\chi^2 = 2\,N \frac{\sigma_p^2}{\bar{p}\,(1-\bar{p})} \;= 2NF_{ST},$$

где F_{ST} – межсубпопуляционный индекс фиксации Райта ($F_{ST} = \sigma_p^2 / \bar{p}(1-\bar{p})$.

Из соотношения для χ^2 следует: $F_{ST} = \chi^2 / 2N$. Для мультиаллельного локуса χ^2-критерий есть (s=2, r>2, df=(s-1)(r-1) = (r-1)):

$$\chi^2 = 2N \left(\sum_k^r \frac{\sigma_{p_k}^2}{\bar{p}_k (1-\bar{p}_k)} \right),$$

где $\bar{p}_k = \sum_i^s w_i\, p_{k_i}$ и $\sigma_{p_k}^2$ – варианса частоты k-ой аллели.

Величина χ^2 для *m* локусов представляет сумму значений χ^2 для каждого локуса и имеет суммарное число степеней свободы. Если по результатам χ^2-теста нулевая гипотеза при критическом уровне статистической значимости $\alpha = 0{,}05$ будет отвергнута ($\chi^2_{\text{факт.}} \geq \chi^2_{\alpha=0{,}05\,\text{табл.}}$), то приступают к расширенному анализу генетического разнообразия выборок.

Пример расчёта χ^2. В табл. 1 даны исходные относительные частоты аллелей и результаты χ^2-теста на соответствие генетических структур выборок.

Расчёты $\bar{p}_i$ и $\sigma_{p_k}^2$ на примере аллели A_3, χ^2-критерия на примере локуса B:

- общая по двум выборкам частота аллели A_3 (аналогично для A_1, A_2, B_1, B_2):

$$\bar{p}_{A_3} = \sum_i^2 w_i p_i = 0{,}294 \times 0{,}10 + 0{,}706 \times 0{,}21 = 0{,}1777,$$

где $w_x = 10/34 = 0{,}294$ и $w_y = 24/34 = 0{,}706$ – «веса» для популяций X и Y;

- межвыборочная варианса ($\sigma^2_{p_k}$) частоты аллели A_3:

$$\sigma^2_{A_3} = \sum_i^2 w_i p_i^2 - \bar{p}^2 =$$

$$= 0{,}294 \times 0{,}10^2 + 0{,}706 \times 0{,}21^2 - 0{,}1777^2 = 0{,}0025;$$

- χ^2-критерий для различий выборок по аллели A_3:

$$\chi^2_{A_3} = 2N \frac{\sigma^2_{A_3}}{\bar{p}_{A_3}} = 2 \times 34 \frac{0{,}0025}{0{,}1777} = 0{,}9613;$$

χ^2-критерий по локусу B (критическое значение $\chi^2_{\alpha=0,05}$=3,84):

$$\chi^2_B = 2N \frac{\sum_i^2 w_i p_i^2 - \bar{p}^2}{\bar{p}(1-\bar{p})} = 2 \times 34 \frac{((0{,}294 \times 0{,}82^2) + (0{,}706 \times 0{,}32^2)) - 0{,}467^2}{0{,}467(1-0{,}467)} = 14{,}1761.$$

χ^2-критерий, суммарный по локусам A и B:

$\chi^2 = 1{,}2857 + 14{,}1761 = 15{,}4618$ при критическом значении $\chi^2_{\alpha=0,05} = 7{,}81$.

Таблица 1. χ^2 ***-тест на соответствие генетических структур выборок X и Y***

Локус/ аллель	Частота, p_i, в выборке		$\bar{p}_i$	$\sigma^2_{p_k}$	χ^2 факт.	df	$\chi^2_{\alpha=0,05}$ табл.
	X (n=10)	Y (n=24)					
Локус A							
A_1	0,47	0,45	0,4559	0,0001	0,0124	-	-
A_2	0,43	0,34	0,3665	0,0017	0,3120	-	-
A_3	0,10	0,21	0,1777	0,0025	0,9613	-	-
	-	-	-	Σ	1,28	2	5,99
Локус B							
B_1	0,82	0,32	0,4670	0,0519	7,55	-	-
B_2	0,18	0,68	0,5330	0,0519	6,62	-	-
Σ	-	-	-	-	14,18	1	3,84
Σ Σ	-	-	-	-	15,46	3	7,81

Использованные данные не позволили выявить различие выборок по частотам аллелей локуса A. По локусу B и двум локусам (A+B) величины χ^2 были выше критических значений при уровне статистической значимости 5% (табл. 1). Следовательно, выборки имели разные профили по аллельным частотам. Это даёт основание приступить к анализу генетического разнообразия по Нею (заметим, дифференциация выборок по Райту была $F_{ST} = \chi^2/2N = 15{,}4618/2 \times 34 = 0{,}2274$ или ≈23%).

Генетические дистанции Нея

Генетическая дистанция – это мера генетических различий между двумя популяциями (видами, породами, линиями, стадами). Исследуя природные сообщества, полагают, что генетическая дистанция зависит от времени, прошедшего с начала дивергенции сравниваемых популяций, имеющих в прошлом общего предка. При этом используют определенную генетическую модель, в которой конкретизируются процессы, приводящие к дивергенции популяций. Например, процессы мутаций генов, случайного дрейфа генов, естественного отбора.

В качестве меры генетических различий двух популяций Ней предложил три дистанции: минимальная, стандартная и максимальная (Nei, 1978; Nei, Roychoudhury, 1974). Эти меры – «вероятности, которые измеряют число замещений генов или кодонов на локус после дивергенции двух рассматриваемых популяций» (Nei, 1987). Поэтому абсолютные величины этих мер имеют чёткое биологическое значение.

Минимальная генетическая дистанция. Допустим, что из популяций X и Y сформированы две выборки, численностью n_X и n_Y, в каждой из которых обследовано по m одних и тех же локусов. Пусть $x_{k\ell}$ и $y_{k\ell}$ – это оценки относительных частот k-ой аллели ℓ-го локуса в выборках X и Y.

Вероятность *идентичности* двух случайно извлечённых генов в популяции X Ней определил как $j_{X_\ell} = \sum_k^{r_\ell} x_{k\ell}^2$, а в популяции Y – $j_{Y_\ell} = \sum_k^{r_\ell} y_{k\ell}^2$ (гомозиготность по Харди-Вайнбергу; вероятность того, что случайно извлечённые особи будут иметь одинаковые аллели). Вероятность идентичности гена, извлечённого из популяции X, и гена, извлечённого из популяции Y, была определена им как $j_{XY_\ell} = \sum_k^{r_\ell} x_{k\ell} y_{k\ell}$ (взаимная идентичность обеих выборок). По всем m локусам (включая мономорфные) это будут усреднённые вероятности: $J_X = (1/m)\sum_\ell^m j_{X_\ell}$, $J_Y = (1/m)\sum_\ell^m j_{Y_\ell}$ и $J_{XY} = (1/m)\sum_\ell^m j_{XY_\ell}$, соответственно. Как считал Ней, определённая таким образом идентичность генов не требует каких-либо допущений о мутации, миграции и отборе. Если последнего нет, и каждая аллель есть производная мутации предшествующих поколений, то ожидаемые значения J_X и J_Y равны коэффициентам инбридинга Райта в популяциях X и Y, а J_{XY} – коэффициенту родства Малекота (Malecot).

Соответствующие *неидентичности* генов были выражены Неем как $D_X = 1 - J_X$, $D_Y = 1 - J_Y$ и $D_{XY} = 1 - J_{XY}$ – все равны доле различных генов (аллелей) между двумя случайно извлечёнными геномами из соответствующих популяций (в частности, D_{XY} отражает пропорцию различных генов между двумя случайно извлечёнными геномами из популяций X и Y). Ней назвал их «минимальные оценки генного разнообразия» (гетерозиготности).

Минимальную генетическую дистанцию (D_{min}) Ней выразил соотношениями:

$$\begin{aligned} D_{min} &= D_{XY} - (D_X + D_Y)/2 = \\ &= (J_X + J_Y)/2 - J_{XY} = \\ &= (1/m)\sum_\ell^m [(j_{X_\ell} + j_{Y_\ell})/2 - j_{XY_\ell}] = \\ &= (1/m)\sum_\ell^m \sum_{k=1}^{r_\ell} (x_{k\ell} - y_{k\ell})^2/2 = \\ &= (1/m)\sum_\ell^m d_\ell , \end{aligned}$$

где $d_\ell = (j_{X_\ell} + j_{Y_\ell})/2 - j_{XY_\ell}$ – оценка генетической дистанции по ℓ-му локусу.

Если выборки небольшие (<50 особей), то вероятности генной идентичности корректируются на численность особей в каждой выборке (Nei, 1978):

$$cJ_X = \frac{1}{m}\sum_\ell^m \frac{2n_X j_{X_\ell} - 1}{2n_X - 1}, \qquad cJ_Y = \frac{1}{m}\sum_\ell^m \frac{2n_Y j_{Y_\ell} - 1}{2n_Y - 1}.$$

Заменив J_X и J_Y на cJ_X и cJ_Y, Ней предложил несмещённую (unbiased) оценку минимальной генетической дистанции (uD_m):

$$uD_{min} = (cJ_X + cJ_Y)/2 - J_{XY}.$$

При низком уровне внутрипопуляционной гомозиготности корректировка может искажать результаты. Ней отмечал: «Недостатком D_{min}, является то, что $D_{X(m)}$, $D_{Y(m)}$ и $D_{XY(m)}$ есть пропорции разных генов в двух случайно извлечённых геномах, так что они не пропорциональны числу различных кодонов. Поэтому, если значение $D_{XY(m)}$ большое, то uD_{min}-статистика может сильно недооценить число чистых кодонных различий» (Nei,1987).

Варианса (V) и стандартная ошибка (SE) uD_{min}, могут быть определены по (Nei, Roychoudhury, 1974):

$$V(uD_{min}) = \frac{\sum_{\ell=1}^m (ud_\ell - uD_{min})^2}{m(m-1)} \quad \text{и} \quad SE(uD_{min}) = \sqrt{V(uD_{min})},$$

где

$$ud_\ell = \frac{1}{2}\left(\frac{2n_X \sum_k x_{k\ell}^2 - 1}{(2n_X - 1)} + \frac{2n_Y \sum_k y_{k\ell}^2 - 1}{(2n_Y - 1)} \right) - \sum_k x_{k\ell} y_{k\ell} .$$

Стандартная генетическая дистанция. По Нею (Nei, 1987), если индивидуальная изменчивость кодонов независима и имеет распределение Пуассона, то среднее число чистых кодонных различий (замещений) между популяциями X и Y может быть выражена через «нормализованную идентичность генов» (генетическое сходство). Как и при расчёте uD_m, первоначало вычисляются J_X, J_Y, J_{XY} и cJ_X, cJ_Y, которые затем участвуют в расчёте нормализованной вероятности того, что две аллели из разных популяций будут идентичными, именно: I_N и её несмещённой оценки uI_N

$$I_N = \frac{J_{XY}}{\sqrt{J_X J_Y}} \quad \text{и} \quad uI_N = \frac{J_{XY}}{\sqrt{cJ_X cJ_Y}}.$$

Величина I_N (uI_N) равна отношению долей идентичных аллелей в разных выборках и в объединённой выборке, т.е. измеряет пропорцию общих аллелей в двух исследуемых выборках (нормированный коэффициент идентичности генов или *индекс генетического сходства*). Диапазон I_N (uI_N) от 0, когда между выборками нет общих аллелей, до 1, когда обе выборки имеют одни и те же аллели с одинаковыми частотами. «Стандартная генетическая дистанция» по Нею вычисляется через натуральный логарифм индекса генетического сходства:

$$D_N = -\ln (I_N).$$

Мотоо Кимура отмечал: «Стандартная генетическая дистанция есть мера среднего числа различий в кодонах. В условиях полной изоляции она даёт суммарное число замен на локус, накопившихся после того, как рассматриваемые популяции дивергировали от общей предковой популяции. Тот факт, что скорость молекулярной эволюции, выраженная числом мутационных замен, примерно постоянна в расчёте на год, делает этот способ измерения генетической дистанции особенно полезным для исследования молекулярной эволюции» (Kimura, 1985; с. 281). Другими словами, D_N – оценка среднего числа замен в каждом локусе, произошедших за время раздельной эволюции двух (суб)популяций. Метод учитывает то обстоятельство, что замены аллелей могут быть неполными: в какой-то части популяции «новый» аллель может вытеснить «старый», который, тем не менее, с большей или меньшей частотой продолжает присутствовать в популяции (Ayala, 1984).

D_N – одна из самых популярных мер генетического разнообразия. Она базируется на допущении, что различия между популяциями обусловлены мутациями и дрейфом генов. Её несмещённая оценка:

$$uD_N = -\ln (uI_N) =$$

$$= -\ln\left(\frac{J_{XY}}{\sqrt{cJ_X\, cJ_Y}}\right) = [(\ln cJ_X + \ln cJ_Y)/2] - \ln J_{XY}.$$

Диапазон значений uD_N – от нуля, при равных частотах аллелей в обеих популяциях, до бесконечности, если в популяциях нет общих аллелей. Последнее связано с тем, что в процессе эволюции, протекающей в течение длительного времени, аллели в каждом локусе могут неоднократно полностью замещаться. Эта оценка адекватна, если темп мутаций аллелей во всех локусах постоянный. Если темпы мутаций из локуса в локус различные, то значение uD_N недооценивается.

Для изолированных родственных популяций $uI_N>0,9$, а $uD_N<0,1$; для дивергирующих – $uI_N<0,8$ и $uD_N>0,2$ (Bader, 1998). В общем, при дивергенции популяций uI_N снижается, а uD_N увеличивается. Следует отметить, что *иногда*, при анализе небольших выборок, значение uD_N может быть больше нуля, даже если две популяции генетически идентичны. Ней (Nei, 1973) назвал это «ложной дистанцией». При небольших размерах выборок значение uD_N может быть отрицательным. В таких случаях uD_N приравнивают к нулю (Nei, 1978).

Приближенные формулы расчёта варианс и стандартных ошибок (Nei, 1987; в развёрнутом виде см. (Zhivotovsky, 1991)):

$$V(uI_N) = uI_N\,(1\text{-}uI_N)/m \quad \text{и} \quad SE(uI_N) = \sqrt{V(uI_N)};$$

$$V(uD_N) = (1\text{-}uI_N)/(uI_N \times m) \quad \text{и} \quad SE(uD_N) = \sqrt{V(uD_N)}.$$

Формулы применимы в тех случаях, когда $uI_N<0,9$ и усреднённая по выборкам гетерозиготность

небольшая.

Здесь целесообразно упомянуть о D_A-дистанции Нея (Nei et al., 1983), которая есть сокращённый (редуцированный) вариант хордовой D_C-дистанции Кавалли-Сфорца и Эдвардса (Cavalli-Sforza, Edwards, 1967):

$$D_A = 1 - (1/m) \sum_{\ell}^{m} \sum_{k=1}^{r_\ell} \sqrt{x_{k\ell} \times y_{k\ell}} \ .$$

Максимальная величина $D_A=1$ достигается тогда, когда две популяции не имеют общих аллелей ни в одном из локусов. D_A-дистанцию рекомендуют использовать на близко родственных популяций, в которых основным фактором генетической дифференциации является генетический дрейф, что часто происходит в случае аборигенных пород домашнего скота (FAO, 2007/2010). Ней считал D_A-дистанцию лучшей мерой для реконструкции филогении *природных* популяций (Takezaki, Nei, 1996) методом «присоединения соседей» (Saitou, Nei, 1987).

Максимальная генетическая дистанция. Выше отмечалось, что если темпы замещения аллелей в разных локусах будут отличаться, то оценка uD_N может быть заниженной. Для этого случая Ней предложил вероятности идентичности генов выражать через среднее геометрическое. В частности, по выборке X – $J'_X = \sqrt[m]{\prod_{\ell=1}^{m} j_{X_\ell}}$, по выборке Y – $J'_Y = \sqrt[m]{\prod_{\ell=1}^{m} j_{Y_\ell}}$, по выборкам X и Y – $J'_{XY} = \sqrt[m]{\prod_{\ell=1}^{m} j_{XY_\ell}}$ (где $\prod$ – символ умножения).

Как и раннее, нормализованная идентичность генов выборок X и Y будет

$$I'_N = \frac{J'_{XY}}{\sqrt{J'_X J'_Y}}$$

а генетическая дистанция:

$$D_{max} = -\ln (I'_N).$$

Величина оценки D_{max} значительно зависит от выборочных ошибок аллельных частот и случайного дрейфа генов. Ней считал, что эти факторы приводят к повышению (инфляции) оценки D_{max}. Поэтому он назвал D_{max} «максимальной генетической дистанцией». Её несмещённая оценка (uD_{max}) рассчитывается по подобию для uD_N:

$$cJ'_X = \sqrt[m]{\prod_{\ell=1}^{m} \frac{2n_X j_{X_\ell} - 1}{2n_X - 1}} \quad \text{и} \quad cJ'_Y = \sqrt[m]{\prod_{\ell=1}^{m} \frac{2n_Y j_{Y_\ell} - 1}{2n_Y - 1}},$$

$$uI'_N = \frac{J'_{XY}}{\sqrt{cJ'_X \, cJ'_Y}} \quad \text{и} \quad uD_{max} = -\ln (uI'_N) \ .$$

Также выборочная варианса и стандартная ошибка uD_{max} рассчитываются по аналогии с таковыми для uD_N.

Вообще, для получения выборочных варианс генетических дистанций наиболее приемлемыми являются методы численного ресэмплинга (имитация взятия новых выборок), в частности, джекнайф-метод (Weir, 1995). Ресэмплинг по локусам имитирует генетические выборки и, следовательно, позволяет получать адекватные оценки варианс для «случайных» популяций. Если D – это оценка любой генетической дистанции по m локусам и D_i – это джекнайф-оценка по i-му локуса, то выборочная варианса есть (Reynolds et al., 1983):

$$V(D) = \frac{m-1}{m} \sum_{i=1}^{m} (D_i - \frac{1}{m} \sum_{j=1}^{m} D_j)^2 \ ,$$

а новая джекнайф-оценка, которая имеет меньшее смещение, чем первоначальная:

$$D^* = m D - \frac{m-1}{m} \sum_{i=1}^{m} D_i \ .$$

Сравнение генетических дистанций. При наличии оценок, например, минимальных генетических дистанций по двум парам выборок (uD_{min1} и uD_{min2}), различие между ними могут быть тестированы следующим образом.

Пусть $uD_{min} = \sum_{\ell}^{m} ud_{\ell} / m$, тогда различие между uD_{min1} и uD_{min2} есть (Nei, 1987):

$$uD_{min_1} - uD_{min_2} = (1/m) \sum_{\ell}^{m} (ud_{\ell_1} - ud_{\ell_2}) =$$
$$= (1/m) \sum_{\ell}^{m} \Delta_{\ell} .$$

где $\Delta_{\ell} = ud_{\ell_1} - ud_{\ell_2}$ – это различие в генетических дистанциях, рассчитанных по ℓ-му локусу; ud_{ℓ} – см. выше. Нулевую гипотезу проверяют, используя обычный двухвыборочный t-критерий Стьюдента. Δ_{ℓ} не имеет нормального распределения, но t-критерий дает приблизительный уровень статистической значимости. Статистически значимое различие между uD_{min1} и uD_{min2} подразумевает таковое между uD_{N1} и uD_{N2}. При наличии джекнайф-оценок двух генетических дистанций, различие между ними считается статистически значимым, если их 95% доверительные (толерантные) интервалы не перекрываются.

Примеры расчёта генетических дистанций. Оценивание генетических дистанций Нея иллюстрируется на двух выборках из гипотетических популяций X и Y. В табл. 2 даны выборочные частоты аллелей и некоторые промежуточные результаты.

Таблица 2. ***Частоты аллелей*** ($x_{k\ell}$ ***и*** $y_{k\ell}$) ***и вероятности идентичности генов*** ($x_{k\ell}^2$, $y_{k\ell}^2$ ***и*** $x_{k\ell}y_{k\ell}$) ***по выборкам X и Y***

Локус, ℓ	Аллель, k	Выборка X (n=10)		Выборка Y (n=24)		$x_{k\ell} \times y_{k\ell}$	d_{ℓ}
		$x_{k\ell}$	$x_{k\ell}^2$	$y_{k\ell}$	$y_{k\ell}^2$		
A	A_1	0,47	0,2209	0,45	0,2025	0,2115	-
	A_2	0,43	0,1849	0,34	0,1156	0,1462	-
	A_3	0,10	0,0100	0,21	0,0441	0,0210	-
	∑	-	0,4160	-	0,3622	0,3787	0,0104
B	B_1	0,82	0,6724	0,32	0,1024	0,2624	-
	B_2	0,18	0,0324	0,68	0,4624	0,1224	-
	∑	-	0,7048	-	0,5648	0,3848	0,2500

Расчёт минимальной генетической дистанции, D_{min}:

$J_X = (1/m) \sum_{\ell}^{m} \sum_{k}^{r_\ell} x_{k\ell}^2 =$ (1/2)(0,4160+0,7048) = 0,5604;

$J_Y = (1/m) \sum_{\ell}^{m} \sum_{k}^{r_\ell} y_{k\ell}^2 =$ (1/2)(0,3622+0,5648) = 0,4635;

$J_{XY} = (1/m) \sum_{\ell}^{m} \sum_{k}^{r_\ell} x_{k\ell} y_{k\ell} =$ (1/2)(0,3787+0,3848) = 0,3818;

$D_{min} = (J_X + J_Y)/2 - J_{XY}$ = (0,5604+0,4635)/2 – 0,3818 = 0,1302.

Расчёт D_{min} по локусам (d_{ℓ}):

$d_A = (j_{X_A} + j_{Y_A})/2 - j_{XY_A} =$ (0,4160+0,3622)/2 – 0,3787 = 0,0104;

$d_B = (j_{X_B} + j_{Y_B})/2 - j_{XY_B} =$ (0,7048+0,5648)/2 – 0,3848 = 0,2500;

$D_{min} = (d_A + d_B)/2$ = (0,0104+0,2500)/2 = 0,1302.

Расчёт вариансы и стандартной ошибки D_m:

$$V(D_{min}) = \frac{\sum_{\ell=1}^{m} (d_{\ell} - D_{min})^2}{m(m-1)} = \frac{(0{,}0104-0{,}1302)^2 + (0{,}2500-0{,}1302)^2}{2(2-1)} = 0{,}0143;$$

$$SE(D_{min}) = \sqrt{V(D_{min})} = \sqrt{0{,}0143} = 0{,}1198.$$

Оценка D_{min}=0,1302 статистически незначимая, т.к. 2×0,1198>0,1302.

Расчёт стандартной генетической дистанции, uD_N:

$I_N = J_{XY} / \sqrt{J_X J_Y} =$ 0,3818/$\sqrt{0{,}5604 \times 0{,}4635}$ = 0,74914.

Оценки J_{XY}, J_X и J_Y взяты из расчёта D_{min}.

$D_N = -\ln(I_N) = -\ln(0{,}74914) = 0{,}2888.$

Несмещенная оценка:

$$cJ_X = \frac{1}{m}\sum_{\ell}^{m}\frac{2n_X\, j_{X_\ell}-1}{2n_X-1} = \frac{1}{2}\left(\frac{2\times10\times0{,}4160-1}{2\times10-1}+\frac{2\times10\times0{,}7048-1}{2\times10-1}\right) = 0{,}5373;$$

$$cJ_Y = \frac{1}{m}\sum_{\ell}^{m}\frac{2n_Y\, j_{Y_\ell}-1}{2n_Y-1} = \frac{1}{2}\left(\frac{2\times24\times0{,}3622-1}{2\times24-1}+\frac{2\times24\times0{,}5648-1}{2\times10-1}\right) = 0{,}4521;$$

$J_{XY} = 0{,}3818;$

$$uI_N = J_{XY}/\sqrt{cJ_X\, cJ_Y} = 0{,}3818/\sqrt{0{,}5373\times0{,}4521} = 0{,}7747;$$

$uD_N = -\ln(uI_N) = -\ln(0{,}7747) = 0{,}2553.$

$uD_N = 0{,}255$ означает, что за время раздельной эволюции двух популяций в каждых 100 локусах в среднем произошло 25,5 аллельных мутаций (замен) или 0,26 замен на один локус.

Отметим: $uD_{min} = (cJ_X + cJ_Y)/2 - J_{XY} = (0{,}5373 + 0{,}4521) - 0{,}3818 = 0{,}1129.$

Варианса и стандартная ошибка uD_N:

$$Var(uD_N) = \frac{(1-uI_N)}{uI_N\times m} = \frac{1-0{,}7747}{0{,}7747\times2} = 0{,}1454;$$

$$SE(uD_N) = \sqrt{Var(uD_N)} = \sqrt{0{,}1454} = 0{,}3813.$$

Расчёт максимальной генетической дистанции, D_{max}:

$$J'_X = \sqrt[m]{\prod_{\ell=1}^{m} j_{X_\ell}} = \sqrt{0{,}4160\times0{,}7048} = 0{,}54148;$$

$$J'_Y = \sqrt[m]{\prod_{\ell=1}^{m} j_{Y_\ell}} = \sqrt{0{,}3622\times0{,}5648} = 0{,}45229;$$

$$J'_{XY} = \sqrt[m]{\prod_{\ell=1}^{m} j_{XY_\ell}} = \sqrt{0{,}3787\times0{,}3848} = 0{,}38174;$$

$$I'_N = J'_{XY}/\sqrt{J'_X\, J'_Y} = 0{,}38174/\sqrt{0{,}54148\times0{,}45229} = 0{,}7714;$$

$$D_{max} = -\ln(I'_N) = -\ln(0{,}7714) = 0{,}2595.$$

Несмещённая оценка:

$$cJ'_X = \sqrt[m]{\prod_{\ell=1}^{m}\frac{2n_x\, j_{X_\ell}-1}{2n_x-1}} = \sqrt[2]{\frac{2\times10\times0{,}4160-1}{2\times10-1}\times\frac{2\times10\times0{,}7048-1}{2\times10-1}} = 0{,}5154;$$

$$cJ'_Y = \sqrt[m]{\prod_{\ell=1}^{m}\frac{2n_Y\, j_{Y_\ell}-1}{2n_Y-1}} = \sqrt[2]{\frac{2\times24\times0{,}3622-1}{2\times24-1}\times\frac{2\times24\times0{,}5648-1}{2\times24-1}} = 0{,}440;$$

$$uI'_N = J'_{XY}/\sqrt{cJ'_X\, cJ'_Y} = 0{,}38174/\sqrt{0{,}5154\times0{,}440} = 0{,}8017;$$

$$uD_{max} = -\ln(uI'_N) = -\ln(0{,}8017) = 0{,}221.$$

Варианса и стандартная ошибка uD_{max}:

$$Var(uD_{max}) = \frac{(1-uI'_N)}{uI'_N\times m} = \frac{1-0{,}8017}{0{,}8017\times2} = 0{,}1237;$$

$$SE(uD_{max}) = \sqrt{Var(uD_{max})} = \sqrt{0{,}1237} = 0{,}3517.$$

Коэффициент генной дифференциации

Базовыми мерами генетического разнообразия популяций являются индексы фиксации Райта (Wright, 1943,1951). Они характеризуют индивидуальный (F_{IS}), субпопуляционный (F_{ST}) и популяционный (F_{IT}) уровни биологической организации подразделённой популяции (детали см. Kuznetsov, 2014). Райт определи F_{ST} для диаллельного локуса, как «корреляцию между двумя аллелями, извлечёнными случайным образом из двух субпопуляций относительно аллелей, извлечённых случайным образом из объединённой популяции». F_{ST}=1, когда обе субпопуляции полностью

гомозиготные и альтернативные аллели фиксированы (отсюда название – «индекс фиксации»), и $F_{ST}=0$, когда частоты аллелей в субпопуляциях одинаковы. В терминах вариансы частот аллелей F_{ST} по Райту:

$$F_{ST}=\frac{V_p}{p(1-p)},$$

где p и $V_p=\sigma_p^2$ – среднее и вариансы частот аллелей по субпопуляциям с диаллельным локусом.

Здесь F_{ST} – это отношение наблюдаемой вариансы, к максимально возможной вариансе при случайном спаривании (гетерозиготность по Харди-Вайнбергу). Райт (Wright, 1978) отмечал, что F_{ST} можно интерпретировать как меру степени дифференциации субпопуляций относительно предельного уровня при полной фиксации (начальная точка движения от исходной гетерозиготности к полной фиксации генов).

Ней (Nei, 1973, 1977) использовал иной подход. Он исходил из аддитивной модели и показал, что генное разнообразие в популяции, как в целом (H_T), может быть разложено на внутри- (H_S) и межсубпопуляционную (D_{ST}) компоненты:

$$H_T=H_S+D_{ST}.$$

Ней выразил вероятность идентичности двух случайно извлечённых из популяции генов как $J=\sum_k x_k^2$, а неидентичности – $H=1-J$ (x_k – частота k-ой аллели в популяции). Вероятность неидентичности, H, – это мера генетического разнообразия (изменчивости) в популяции, т.е. гетерозиготности. Однако Ней считал, что термин «гетерозиготность» не корректен для популяции, в которой особи спариваются *неслучайно*. Поэтому, он предложил для H использовать словосочетание «gene diversity» (генное разнообразие), а для J – «gene identity» (генная идентичность). При случайном спаривании термины «генное разнообразие» и «генная идентичность» становятся эквивалентными терминам «гетерозиготность» и «гомозиготность», соответственно.

Допустим популяцию, которая подразделена на s субпопуляций и $x_{ik\ell}$ – это частота k-ой аллели (k=1, 2, …, r), относящейся к ℓ-му локусу (ℓ=1, 2, …, m) в i-ой субпопуляции (i=1, 2, …, s). Тогда по ℓ-му локусу имеем вероятности идентичности генов (Nei, 1973):

- в i-ой субпопуляции
$$J_{i.\ell}=\sum_{k=1}^{r} x_{ik\ell}^2;$$
- усреднённой по s субпопуляциям
$$J_{S\ell}=\sum_{i=1}^{s} J_{i.\ell}/s;$$
- во всей популяции (субпопуляции объединены)
$$J_{T\ell}=\sum_{k=1}^{r} \bar{x}_{.k\ell}^2, \quad \text{где } \bar{x}_{.k\ell}=\sum_{i=1}^{s} x_{ik\ell}/s.$$

По оценкам вероятностей идентичности генов получаем ожидаемое генное разнообразие по ℓ-му локусу:

- внутри субпопуляций
$$H_{S\ell}=1-J_{S\ell};$$
- в объединённых субпопуляциях (усреднённое)
$$H_{T\ell}=1-J_{T\ell};$$
- между субпопуляциями (включает сравнение субпопуляции с собой):
$$D_{ST\ell}=H_{T\ell}-H_{S\ell}.$$

$D_{ST\ell}$ – это *абсолютная* мера межсубпопуляционного разнообразия генов (гетерозиготности). Исходя из аддитивной модели, $H_T=H_S+D_{ST}$, *относительная* мера межсубпопуляционного разнообразия генов по ℓ-му локусу есть

$$G_{ST\ell}=\frac{D_{ST\ell}}{H_{T\ell}}=\frac{H_{T\ell}-H_{S\ell}}{H_{T\ell}}=1-\frac{H_{S\ell}}{H_{T\ell}}.$$

Статистика G_{ST} эквивалентна индексу фиксации, F_{ST}, Райта (если локус диаллельный, то было показано, что $H_T = 2\bar{x}(1-\bar{x})$ и $D_{ST} = 2\sigma^2_{\bar{x}}$; в случае множественных аллелей, G_{ST} эквивалентно средневзвешенному F_{ST} по всем аллелям). Ней назвал G_{ST}-статистику «*коэффициентом генной дифференциации*» (coefficient of gene differentiation). G_{ST} интерпретируют как разнообразие генов между субпопуляциями, а отношение H_S/H_T (или $1\text{-}G_{ST}$) – как разнообразие генов внутри субпопуляций.

Величина G_{ST} зависит от анализируемой популяции. Поэтому Ней считал, что оценку, полученную в одной подразделённой популяции нельзя сравнивать с таковой в другой (исключая случаи, когда системы спаривания аналогичны в обеих популяциях).

Из уравнений $H_T = H_S + D_{ST}$ и $G_{ST} = D_{ST}/H_T$ Ней (Nei, 1973) вывел отношение:

$$1\text{-}J_S = (1\text{-}G_{ST})(1\text{-}J_T),$$

подобное известному отношению Райта: $1\text{-}F_{IT} = (1\text{-}F_{IS})(1\text{-}F_{ST})$ (Wright ,1951, 1978). Различие в том, что в последнем F_{IS} и F_{IT} измеряют отклонения наблюдаемых генотипических частот от таковых при равновесии Харди-Вайнберга, в то время как в уравнении Нея J_S и J_T – это вероятности идентичности генов на разных уровнях популяционной структуры. Кроме того, G_{ST}, J_T и J_S всегда положительные.

Как отмечалось выше, D_{ST_ℓ} включает и сравнение субпопуляций с собой. С поправкой на эти сравнения, межсубпопуляционное разнообразие генов есть

$$\overline{D}_{m_\ell} = \frac{s}{s-1} D_{ST_\ell}.$$

Это абсолютная мера генной дифференциации, которая не зависит от разнообразия генов внутри субпопуляций (Nei, 1973). $\overline{D}_{m_\ell}$ используют: (а) для сравнения степени генной дифференциации различных организмов и (б) при расчёте доли межсубпопуляционного генного разнообразия относительно внутрипопуляционного генного разнообразие:

$$R_{ST_\ell} = \overline{D}_{m_\ell} / H_{S_\ell}.$$

Несмещённая оценка **G_{ST} (uG_{ST})**. Если исследуемые выборки небольшого размера, то Ней и Чессер (Nei, Chesser, 1983) ввёл корректировку оценок H_{S_ℓ} и H_{T_ℓ} на выборочную ошибку:

$$cH_{S_\ell} = \frac{\tilde{n}}{\tilde{n}-1}\left(H_{S_\ell} - \frac{H_{O_\ell}}{2\tilde{n}}\right) \quad \text{и}$$

$$cH_{T_\ell} = H_{T_\ell} + \frac{cH_{S_\ell}}{\tilde{n} \times s} - \frac{H_{O_\ell}}{2 \times \tilde{n} \times s},$$

где $\tilde{n} = 1/[(1/s)\sum_i^s 1/n_i]$ – средний (гармоничный) размер выборки; H_{O_ℓ} – усреднённая по популяциям наблюдаемая гетерозиготность по ℓ-ому локусу:

$$H_{O_\ell} = 1 - \sum_i^s X_{ikk\ell}/s,$$

где $X_{ikk\ell}$ – число индивидов с гомозиготными генотипами (kk) в ℓ-ом локусе и i-ой популяции.

Несмещённая оценка G_{ST_ℓ} есть

$$uG_{ST_\ell} = \frac{cH_{T_\ell} - cH_{S_\ell}}{cH_{T_\ell}} = \frac{uD_{ST_\ell}}{cH_{T_\ell}}.$$

Обобщение по локусам. Рассмотренное выше касалось только одного локуса. Метод применим к любому числу локусов. Для этого значения cH_{S_ℓ} и cH_{T_ℓ} усредняются по всем исследуемым локусам (возможно решение через усреднение сначала H_S, H_T и H_O):

$$cH_S = \sum_\ell^m cH_{S_\ell}/m, \quad cH_T = \sum_\ell^m cH_{T_\ell}/m$$

и сводная несмещённая оценка коэффициента относительной генной дифференциации есть

$$uG_{ST} = \frac{cH_T - cH_S}{cH_T} = \frac{uD_{ST}}{cH_T} = 1 - \frac{cH_S}{cH_T}.$$

При небольшом размере выборок uG_{ST} *иногда* может быть отрицательной. При очень большой генной дифференциации популяций величина J_T может быть ниже, чем J_S (гены из разных популяций более схожи, чем гены из одной популяции). Тогда значение D_{ST} (соответственно, $\overline{D}_m$) будет отрицательным. Для таких случаев Ней предложил использовать логарифмы (Nei, 1973):

$$D_{ST} = -\ln (J_T/J_S) \quad \text{и} \quad G_{ST} = -\ln (J_T/J_S) / [-\ln J_T]$$

по аналогии с генетической дистанцией (здесь J = 1-H). Для получения объективных оценок генной дифференциации субпопуляций, выборки должны быть случайно отобранными из популяций и включать большое число полиморфных и мономорфных локусов.

Пример расчёта коэффициента генной дифференциации. Имеются данные по частотам генотипов (табл. 3) в выборках из трех субпопуляций (один локус с тремя аллелями: A, B и C).

Таблица 3. ***Частоты генотипов и оценки наблюдаемой гетерозиготности***

Выборка	n_i	Частота генотипа					H_O
		AA	AB	AC	BB	BC	
1	49	0,2041	0,3265	-	0,4286	0,0408	0,3673
2	82	0,1342	0,4024	0,0122	0,4268	0,0244	0,4390
3	37	0,0270	0,2433	-	0,7027	0,0270	0,2703
Среднее	168	0,1218	0,3241	0,0040	0,5194	0,0307	0,3588

Частота генотипа, например, AA, по объединённой выборке:

$$X_{AA} = (0{,}2041 + 0{,}1342 + 0{,}0270)/3 = 0{,}1218.$$

Аналогично рассчитываются частоты генотипов AB, AC и др.

Расчёт наблюдаемой гетерозиготности в выборке 1:

$$H_{O_1} = 1 - \sum\nolimits_{kk} X_{kk} = 1 - (0{,}2041 + 0{,}4286) = 0{,}3673.$$

Три варианта расчёта наблюдаемой гетерозиготности (H_O) в популяции в целом:

$$H_O = 1 - (0{,}2041 + 0{,}4286 + 0{,}1342 + 0{,}4268 + 0{,}0270 + 0{,}7027)/3 =$$
$$= (0{,}3673 + 0{,}4390 + 0{,}2703)/3 = \quad 1 - (0{,}1218 + 0{,}5194) = 0{,}3588.$$

В табл. 4 приведены частоты аллелей и некоторые промежуточные величины, необходимые для расчёта uG_{ST}-статистики.

Таблица 4. ***Частоты генов и промежуточные величины для расчёта uG_{ST}-статистики***

Выборка	Частота аллели x_{ik}			$J_S = \sum x_k^2$	$H_S = 1 - J_s$
	A	B	C		
1	0,3674	0,6122	0,0204	0,5102	0,4898
2	0,3415	0,6402	0,0183	0,5268	0,4732
3	0,1487	0,8378	0,0135	0,7242	0,2758
$\overline{x}_{.k}$	0,2859	0,6967	0,0174	-	-
$\overline{x_k^2}$	0,0817	0,4955	0,0003	0,5871	0,4129
J_T		0,5674		-	-
H_T		0,4326		-	-

Вероятность идентичности генов:

- в выборке 1

$$J_1 = 0{,}3674^2 + 0{,}6122^2 + 0{,}0204^2 = 0{,}5102;$$

- усреднённой по выборкам

$$J_S = (0{,}5102+0{,}5268+0{,}7242)/3=0{,}5871;$$

- в объединенной выборке

$$J_T = 0{,}2859^2+0{,}6967^2+0{,}0174^2 =$$
$$= 0{,}0912+0{,}4955+0{,}0003 = 0{,}5674.$$

Генное разнообразие:

- усреднённое по выборкам

$$H_S = 1\text{-}J_S = 1\text{-}0{,}5871 = 0{,}4129;$$

- в объединённой выборке

$$H_T = 1\text{-}J_T = 1\text{-}0{,}5674 = 0{,}4326.$$

Абсолютное генное разнообразие между выборками:

$$D_{ST} = H_T\text{-}H_S = 0{,}4326 - 0{,}4129 = 0{,}0197.$$

$D_{ST,}$ скорректированное по числу выборок:

$$\overline{D}_m = (s/(s\text{-}1)\ D_{ST} = [3/(3\text{-}1)]\ 0{,}0197 = 0{,}0296.$$

Отношение генного разнообразия между выборками к усреднённому генному разнообразию по выборкам:

$$R_{ST} = \overline{D}_m/H_S = 0{,}0296/0{,}4129 = 0{,}0717.$$

Коэффициент генной дифференциации выборок:

$$G_{ST} = D_{ST}/H_T = 0{,}0197/0{,}4326 = 0{,}0455.$$

Проверка соотношения $1\text{-}J_S = (1\text{-}G_{ST})(1\text{-}J_T)$:

$$1\text{-}0{,}5871 = (1\text{-}0{,}0455)(\ 1\text{-}0{,}5674)$$
$$0{,}4129 = 0{,}9545\times0{,}4326$$
$$0{,}4129 = 0{,}41292.$$

Средний (гармоничный) размер выборок:

$$\tilde{n} = \frac{1}{(1/s)\sum_i^s 1/n_i} = \frac{1}{\frac{1}{3}\left(\frac{1}{49}+\frac{1}{82}+\frac{1}{37}\right)} = 50{,}3\cdot$$

Корректировка H_S и H_T для получения несмещённой оценки G_{ST}:

$$cH_S = \frac{\tilde{n}}{\tilde{n}-1}(H_S - \frac{H_O}{2\tilde{n}}) = \frac{50{,}3}{50{,}3-1}\left[0{,}4129 - \frac{0{,}3588}{2\times50{,}3}\right] = 0{,}4175;$$

$$cH_T = H_T + \frac{cH_S}{\tilde{n}\times s} - \frac{H_O}{2\times\tilde{n}\times s} = 0{,}4326 + \frac{0{,}4175}{50{,}3\times3} - \frac{0{,}3588}{2\times50{,}3\times3} = 0{,}4342.$$

Несмещенная оценка (uG_{ST}):

$$uG_{ST} = \frac{0{,}4342-0{,}4175}{0{,}4342} = 0{,}0384 \text{ или } \approx 3{,}8\%.$$

Математическое ожидание для D_N, D_{min} и G_{ST}

Анализ генетического разнообразия природных видов проводится не столько для выявления степени генетической дифференциации популяций, сколько для оценки таких демографических параметров, как размер эффективной популяции, степень генного потока, период времени дивергенции, а также для реконструкции филогенетических отношений (суб)популяций и построения дендрограммы. При этом принимается та или иная популяционно-генетическая модель. Так, в модели «равновесной (equilibrium) популяции» считается, что эффекты действия разных эволюционных сил, таких как мутация, миграция, генетический дрейф и естественный отбор, находятся в равновесии, так что частоты генов (аллелей) в популяции остаются неизменными.

«Островная модель» Райта (island model) допускает, что популяция разделена на *s* субпопуляций, каждая из которых имеет эффективную численность n_e, находится в состоянии равновесия по Харди-Вайнбергу (индивиды размножаются случайным образом) и с равным шансом может включать иммигрантов (доля m) из других субпопуляций (обмен генами между субпопуляциями происходит с одинаковой скоростью).

Мутационная модель «бесконечного числа аллелей» (infinite alleles model, IAM) предполагает, что: (а) каждая мутация приводит к появлению новой, не существовавшей ранее в популяции, аллели с заданной скоростью μ; (б) предковая популяция находилась в состоянии равновесия по Харди-Вайнбергу; (в) разделение предковой популяции на субпопуляции X и Y было моментальным и полным; (г) субпопуляции X и Y полностью изолированы, имеют постоянный эффективный размер, равный эффективному размеру предковой популяции; (д) вероятности идентичности генов в субпопуляциях X и Y равны таковой в предковой популяции.

В модели «пошаговой (ступенчатой) мутации» (stepwise mutation model**,** SMM) каждая мутация создает новый аллель, добавляя или удаляя повторный мотив с равной вероятностью μ/2 в обоих вариантах. Иногда допускают, что нет никаких ограничений на число повторов, возможных в локусе. Следовательно, аллели очень разных размеров будут более отдаленно связаны, чем аллели аналогичных размеров. Считается, что SMM имеет «память» размера аллели и более точно отражает процесс мутации микросателлитов.

В случае IAM и отсутствия дифференцированного отбора на протяжении всего эволюционного процесса, вероятность идентичности двух генов (из двух субпопуляций) будет уменьшаться со скоростью 2μ на поколение. Тогда математическое ожидание* идентичности в t-ом поколении есть (Kimura, 1985)

$$E(I_t) = (1-2\mu)\, E(I_{t-1}) = (1-2\mu)^t I_0 \approx e^{-2\mu t} I_0.$$

где I_0 – вероятность идентичности (гомозиготности) в предковой популяции (t=0); μ – темп мутирования на локус для нейтральных аллелей (равный для всех локусов); t – число поколений, прошедших с начала дивергенции.

Это значит, что математическое ожидание для стандартной генетической дистанции есть

$$E(D_N) \approx 2\mu t.$$

То есть, эта величина увеличивается во времени пропорционально (линейно) темпу мутаций. Поэтому D_N применяют в случае длительной эволюции, при которой субпопуляции дивергируют в результате мутаций и дрейфа генов. Для минимальной генетической дистанции, D_{min}, математическое ожидание есть (Takezaki, Nei, 1996)

$$E(D_{min}) = J(1 - e^{-2\mu t}),$$

где J – ожидаемая гомозиготность двух популяций.

Из математического ожидания для D_N можно получить время расхождения субпопуляций:

$$t \approx D_N/2\mu.$$

Если изоляция между субпопуляциями будет не абсолютной (что характерно для многих реальных локальных популяций), то миграция (поток генов) будет препятствовать процессу дивергенции. Это отразится негативно на величине генетической дистанции и, следовательно, на оценке времени расхождения популяций. Была предложена меры времени дивергенции по микросателлитным локусам, которая не зависела от динамики популяции и достаточно устойчива к слабому потоку генов (Zhivotovsky, 2001, 2006).

В очень представительном исследовании (Gautier et al., 2007) были получены генетические дистанции между девятью европейскими породами крупного рогатого скота, которые в среднем составили 0,04±0,01, между шестью африканскими породами – 0,04±0,03 и между европейскими и

---

* Термин «математическое ожидание» связан с начальным периодом возникновения теории вероятности, когда область её применения ограничивалась азартными играми. Игрока интересовало среднее значение ожидаемого выигрыша или, иначе, математическое ожидание выигрыша. Для математического ожидания случайной величины X используют обозначения E(X) или M(X) (Кремер Н.Ш. Теория вероятности и математическая статистика. – М.: ЮНИТИ-ДАНА, 2006, 573 с.).

африканскими породами – 0,11±0,02. По (MacHugh, 1996) темпы мутаций $10^{-5}...10^{-2}$ на ген/поколение. Тогда продолжительность дивергенции оценивается в

$$t \approx 0{,}11/(2\times 0{,}00001) \approx 5500 \text{ поколений.}$$

Если принятые допущения и предпосылки были релевантными (соответствовали действительности), то можно полагать, что разделение предковой популяции на африканскую и европейскую части началось 27-33 тыс. лет назад (при генерационном интервале 5-6 лет) и примерно 10 тыс. лет назад (=(0,04×5)/(2×0,00001)) началась дивергенция внутри этих двух субпопуляций (доместикация).

Для «островной модели» ожидаемое равновесное значение F_{ST} есть (Wright, 1943)

$$F_{ST} \approx \frac{1}{4n_e m + 4n_e \mu + 1}.$$

Из этого соотношения следует, что на степень дивергенции субпопуляций определяющее влияние оказывают не n_e и m, а их произведение, $n_e \times m$, т.е. число действительных иммигрантов за поколение – M.

Леттер (Latter, 1973) предложил соотношение для G_{ST}:

$$G_{ST} \approx \frac{1}{4n_e\left(\frac{s}{s-1}\right)(m+\mu)+1}.$$

При $m >> \mu$ (существенно больше) оно показывает, что G_{ST} зависит только от абсолютного числа мигрантов Nm и числа субпопуляций (s). При $s=\infty$ это выражение тождественно с таковым для F_{ST}. Если допустить, что $\mu << m$, т.е. $4n_e \mu = 0$, то

$$M \approx \frac{1-F_{ST}}{4F_{ST}} \quad \text{и} \quad M \approx \frac{1-G_{ST}}{4G_{ST}}.$$

Для скорости миграции (m) Джост (Jost et al., 2018) привёл следующее соотношение:

$$m = \frac{1/G_{ST} - 1 - 4n_e \mu (s/(s-1))}{4n_e (s/(s-1))^2}.$$

При анализе микросателлитных локусов значения статистик F_{ST} и G_{ST} ограничены величиной гетерозиготности, H_S, в пределах субпопуляций. Хедрик (Hedrick, 2005) предложил стандартизированную меру дифференциации, на основе оригинальной оценки G_{ST}:

$$G'_{ST} = \frac{G_{ST}(s-1+H_S)}{(s-1)(1-H_S)}.$$

Эта стандартизация распространяется и на F_{ST}. Хотя нет прямой связи между стандартизированными статистиками и m, предполагается, что оценка числа мигрантов без влияния (воздействия) H_S может быть получена с помощью комбинации статистик до- и после стандартизации (Meirmans, Hedrick, 2011):

$$M \approx \frac{1-F'_{ST}}{4F_{ST}} \quad \text{и} \quad M \approx \frac{1-G'_{ST}}{4G_{ST}}.$$

Величина M может в определённой степени характеризовать интенсивность потока генов между субпопуляциями. Чем больше степень генной дифференциации субпопуляций, тем меньше величина M. Если $2M<1$, то субпопуляции имеют тенденцию к дивергенции; при $2M>1$ тенденции к дивергенции нет (Holsinger, 2010).

Показано (Hedrick, 2005), что в случае, когда имеют место мутации и миграция и допускается, что $4n_e\mu$ может характеризоваться соотношением $H_S/(1-H_S)$, то

$$M \approx \frac{1-F_{ST}(1+H_S/(1-H_S)}{4F_{ST}}.$$

При фиксированном значении F_{ST} (G_{ST}) повышение H_S (гетерозиготности, т.е. повышение темпа мутации) приводит к снижению ожидаемого числа мигрантов.

Теоретически, ожидаемая гетерозиготность H, есть функция эффективного размера популяции (n_e) и темпа мутации за поколение (μ). В случае IAM и небольшого числа аллельных состояний гена (Takezaki, Nei, 1996)

$$H = \frac{4n_e\mu}{1+4n_e\mu},$$

а для SMM ожидаемая гетерозиготность есть

$$H = 1 - \frac{1}{\sqrt{1+8n_e\mu}}.$$

Ней и Такезаки (Nei, Takezaki, 1994) исследовали *in silico* (компьютерное моделирование) девять мер генетической дистанции и два подхода к построению филогенетического древа (дендрограммы), именно: «метод невзвешенного попарного центроидного усреднения» (unweighted pair-group method with arithmetic means, UPGMA) и метод «присоединения соседей» (neighbor-joining, NJ). В методе UPGMA предполагается, что темпы эволюции одинаковы для всех эволюционных ветвей; NJ-метод не требует такого допущения.

После генерирования данных по частотам аллелей и получения генетических дистанций для каждого набора локусов (10, 20, ..., 100 локусов) с двумя уровнями средней гетерозиготности (H=0,16 и H=0,5), строились филогенетические древа. Их топология сравнивалась с таковой модельного древа. Процедура повторялась 100 раз с расчётом процента повторов правильной топологии (P_C). Было показано, что стандартная генетическая дистанция Нея (D_N) лучше подходит для оценки эволюционного времени, чем D_A. Последняя также может быть использована для этой цели, если рассматривается короткий эволюционный период. При реконструкции филогенетических древ более эффективной была D_A-дистанция (относительно D_N). При уровне гетерозиготности H=0,16 коэффициент P_C был всегда выше для NJ-метода, чем для UPGMA. Исключением была D_N-дистанция, когда метод UPGMA показывал более высокие значения P_C. Однако при H=0,5 NJ-метод давал более низкие значения P_C, чем UPGMA почти со всеми мерами генетических дистанций. В аналогичной работе (Takezaki, Nei, 1996) D_A-дистанция Нея и хордовая D_C-дистанция Кавалли-Сфорца и Эдвардса (Cavalli-Sforza, Edwards, 1967) как для IAM, так и для SMM, показывали, как правило, более высокие значения P_C, чем при использовании других мер. Результаты не зависели от того, имел или не имел место эффект «бутылочного горлышка» (сужения и расширения численности популяции).

Заключительные замечания

Имеется достаточно много индексов фиксации, мер дистанции и коэффициентов дифференциации, так или иначе характеризующих генетические различия/сходства между популяциями (Nei et al., 1983; Takezaki, Nei, 1996). Несмотря на то, что эти меры часто базируются на разных биологических и математических предпосылках, их оценки положительно коррелируют (0,74-0,99 по Barker, 1999), особенно при небольших генетических различиях между популяциями. При значительной дифференциации популяций могут иметь место существенные расхождения оценок, полученных разными методами. Это особенно относится к анализу микросателлитных маркеров, когда предполагаются разные типы мутаций (Takezaki, Nei, 1996; Goldstein, Pollock. 1997; Paetkau et al., 1997).

Из-за разных темпов мутаций в локусах и дрейфа генов оценки генетических дистанций между популяциями варьируют по локусам. Окончательная оценка генетической дистанции является функцией частных генетических дистанций по каждому локусу. Если имеется *m* локусов по *n* особям каждого локуса, то выборочная варианса оценки генетической дистанции ($V_{sum(m,n)}$) включает два компонента (Nei, Roychoudhury, 1974; Li, Nei, 1975; Nei 1978; Kalinowski, 2002):

$$V_{sum(m,n)} = V_{inter(m)} + V_{intra(m,n)},$$

где $V_{inter(m)}$ – межлокусная варианса распределения генетических дистанций по локусам; $V_{intra(m,n)}$ – внутрилокусная варианса, которая представляет собой выборочную варианту, относящуюся к выборке ограниченного числа особей по исследуемым локусам.

Межлокусная выборочная варианса сокращается за счёт привлечения большого числа локусов, внутрилокусная варианса – путем отбора большого числа особей. Если межлокусная варианса существенно меньше внутрилокусной, то привлечение большого числа локусов не будет эффективным

способом снижения общей выборочной вариансы, и наоборот. Поэтому для получения статистически значимых оценок генетических дистанций (коэффициентов дифференциаций) и объективных выводов о генетической структуре популяций, выборки локусов и особей должны быть, по-возможности, рандомизированными и достаточно большими. Например, для предварительного анализа по микросателлитам FAO (2007/2010) рекомендует использовать минимум 25 животных, генотипированных по 30 локусами. Лучше, если заранее будет выбрана наиболее эффективная схема исследования (минимальная ошибка результата при минимальных издержках), что предполагает знание относительных величин меж- и внутрилокусных вариант и компьютерное моделирование. В противном случае оценки параметров могут иметь значительные выборочные ошибки.

Здесь уместно упомянуть о биологической и практической значимости оценок генетической дифференциации. Так, в работе (Smaragdov, 2018) F_{ST}-оценки генетических дистанций по SNP-маркерам, равные, например, 0,009, 0,005 и даже 0,002, были статистически значимыми (стандартные ошибки ≤ 0,0003). Однако их биологическая и, тем более, практическая значимость, на наш взгляд, не очевидны, как и утверждение автора, что «коровы в каждом стаде генетически отличаются от животных в других стадах» (они, в принципе, отличаются и в пределах стада). Заметим, что по классификации Райта при F_{ST} < 0,05 (или 5%) генетическая дифференциация популяций считается *незначительной* (Wright,1978). Как нам представляется, дифференциацию менее 1% можно отнести к категории «не имеющий существенного значения или ничтожной» (при F_{ST}=0 субпопуляции не различаются по числу и частотам аллелей). Гомогенность (однородность) стад на уровне 99,1-99,8% означает, что их аллельные профили практически одинаковы. Поэтому интерпретировать подобную «дифференциацию» необходимо с осторожностью в соответствии со здравым смыслом. Стивен Калиновски, который рассматривал *in silico* эволюционные и статистические свойства трёх генетических дистанций, отмечал: «Какая бы генетическая дистанция ни использовалась для обобщения генетических различий между популяциями, самой большой проблемой будет решение, какие эволюционные процессы создали наблюдаемый паттерн, и оценить, какую биологическую значимость это имеет для популяций» (Kalinowski, 2002).

Биологическая основа статистик разнообразия – динамичный эволюционный процесс. Как правило, все меры сходства/различия популяций разрабатывались для того, чтобы исследовать филогенетические отношения между формами жизни через оценку генетических (эволюционных) дистанций. Эти меры базируются на различных моделях «истории эволюции», с разными генетическими и демографических допущениями и предположениями (равновесная популяция, островная модель, IAM, SMM). Даже в природных популяциях с естественных ходом эволюции выполнение всех условий и допущений может быть маловероятным. По мнению Л.А. Животовского (Zhivotovsky, 2006), для популяций человека ни одно из теоретических допущений не выполняется. Животовский (Zhivotovsky, 2001) также показал, что если численность популяции увеличивается, то оценка времени дивергенции занижается. Кроме того, в реальных условиях трудно определить, находятся ли популяции в состоянии равновесия, или нет. Учитывая временной масштаб, необходимый для достижения популяциями равновесия, вполне вероятно, что многие виды далеки от его достижения по многим генетическим локусам (т.е. они всё еще находятся в неравновесном состоянии).

У доместицированных видов животных естественный ход развития нарушается из-за использования разных методов селекции и кроссбридинга, как для улучшения «старых», так и для создания «новых» пород. «Эволюционные истории» пород переплетаются и смешиваются. Теоретические модели, заложенные в статистиках разнообразия, в условиях искусственного разведения животных представляются ещё более нереалистичными. Поэтому эволюционно-демографические показатели и реконструированные иерархические филогенетические деревья, установленные по оценкам генетических дистанций, могут быть далеки от истинных. Как отмечалось в материалах ФАО «Состояние всемирных генетических ресурсов…» (2007/2010), недостаток реконструкции филогенетического древа домашних животных в том, что эволюция его ветвей не может образовывать сеть; ветви могут расходиться, но не могут появляться за счет пересечения. Новые породы, как правило, возникают в результате различных типов скрещивания имеющихся пород. Сложные «эволюционные сценарии» плохо описываются такими методами, как UPGMA и NJ. Поэтому полученные с их помощью «реконструкции» следует воспринимать «с осторожным скептицизмом» (в лучшем случае, как вариант

кластеризации). Возможно, более реалистичными и полезными для теории и практики разведения продуктивных животных будут «филогенетические сети» (phylogenetic networks), предназначение которых анализ и визуализация таких перекрёстных событий, как гибридизация, горизонтальный перенос генов, рекомбинация и т.п. (Huson, Bryant, 2006).

Тем не менее, если задачей исследования является анализ только *текущего* генетического разнообразия конкретных выборок (пород, линий, стад, групп животных), то статистики Нея могут быть хорошей мерой их дифференциации (Holsinger, Weir, 2009). Брюс Вейр (Weir, 2012) отмечал: «Как описание текущих частот, подход Нея является подходящим, но тогда нет никаких оснований для эволюционной интерпретации оценок, и нет никаких оснований для того, чтобы делать заявления о дивергенции от предковых популяций, эффектах естественного отбора или степени миграции». Оценки uD_N и uG_{ST} могут служить ценной дополнительной информацией, позволяющей селекционерам в совокупности с традиционными и биометрическими методами принимать более правильные решения по разведению продуктивных животных. В частности, на основе матриц попарных генетических дистанций породных выборок и их 2D (3D) визуализации (например, методом главных координат – principal coordinate analysis, PCoA) можно генетически обосновать: (а) выбор породы для улучшения стада или иной породы, (б) выбор пород для промышленного скрещивании с тем, чтобы получить эффект гетерозиса, (в) схем линейного разведения, кросса линий и группового подбора для минимизации или оптимизации коэффициента инбридинга, (г) выбор «местных» пород(ы), которые следует сохранить (или поглатить), при условии недопущения снижения генетического разнообразия.

REFERENCES

1. Ayala F.J. *Vvedenie v populyatsionnuyu i evolyutsionnuyu genetiku* (Introduction to population and evolutionary genetics). Moscow: Mir, 1984, 232 p. (In Russian)
2. Bader J.M. Measuring genetic variability in natural populations by allozyme electrophoresis. In: *Tested studies for laboratory teaching, Proceedings of the 19th Workshop/Conference of the Association for Biology Laboratory Education (ABLE).* University of Calgary Alberta, Canada, 1998, 19: 25-41.
3. Barker J.S.F. Conservation of livestock breed diversity. *Animal Genetic Resources Information*. 1999, 25: 33-43.
4. Cavalli-Sforza L.L, Edwards A.W.F Phylogenetic analysis: models and estimation procedures. *Evolution*. 1967, 21(3/1): 550-570.
5. Gautier M., Faraut T., Moazami-Goudarzi K., Navratil V., Foglio M., Grohs C., Boland A., Garnier J.-G., Boichard D., Goldstein D.B., Pollock D.D. Launching microsatellites: A review of mutation processes and methods of phylogenetic inference. *J. Heredity*. 1997, 88(5): 335-342.
6. Holsinger K.E. *Lecture notes in population genetics*. University of Connecticut, 2010, 275 p.
7. Holsinger K.E., Weir B.S. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat. Rev. Genet*. 2009, 10(9): 639–650. DOI:10.1038/nrg2611.
8. Huson D.H., Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol*. 2006, 23(2): 254-267. DOI:10.1093/molbev/msj030.
9. Hedrick P.W. A standardized genetic differentiation measure. *Evolution.* 2005, 59(8): 1633-1638.
10. Hedrik P. *Genetika populyatsii* (Genetics of populations). Moscow: Tekhnosfera Publ., 2003, 592 p. (In Russian)
11. Rischkowsky B., Pilling D. (Eds). *The State of the World's Animal Genetic Resources for Food and Agriculture.* Rome: FAO Publ., 2007 (*Sostoyanie vsemirnykh geneticheskikh resursov zhivotnykh v sfere prodovol'stviya i sel'skogo khozyaistva* (Translation to Russian, Moscow: VIZ Publ., 2010).
12. Smaragdov M.G. [Full genome assessment of cross-breeding genetic differences in cattle]. *Dostizheniya nauki i tekhniki APK - Scientific and Technological Agribusiness.* 2018, 32(4): 47-49. DOI: 10.24411/0235-2451-2018-10411.
13. Jost L., Archer F., Flanagan S., Gaggiotti O., Hoban S., Latch E. Differentiation measures for conservation genetics. *Evol. Appl*. 2018, 11(7, Suppl): 1139-1148. DOI:10.1111/eva.12590.
14. Kalinowski S.T. Evolutionary and statistical properties of three genetic distances. *Mol. Ecol*. 2002, 11(8): 1263-1273.
15. Kimura M. *Molekulyarnaya evolyutsiya: teoriya neitral'nsti* (Molecular evolution: the theory of neutrality) Moscow: Mir, 1985, 394 p. (in Russian)
16. Kuznetsov V.M. [Wright's F-statistics: estimation and interpretation]. *Problemy biologii productivnykh zhivotnykh – Problems of Productive Animal Biology*. 2014, 4: 80-104 (in Russian).
17. Latter B.D.H. The island model of population differentiation: a general solution. *Genetics.* 1973, 73(1): 147-157.
18. Lathrop G.M., Gut I.G., Eggen A. Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics.* 2007, 177(1): 1059-1070. DOI.org/10.1534/genetics.107.075804.
19. Li W., Nei M. Drift variances of heterozygosity and genetic distance in transient states. *Genetics Research Camb*.

1975, 25(3): 229-248.
20. MacHugh D.E. *Molecular biogeography and genetic structure of domesticated cattle*. A thesis submitted for the degree of Doctor of Philosophy, Trinity College, University of Dublin, 1996, 264 p.
21. Meirmans P.G., Hedrick P.W. Assessing population structure: F_{ST} and related measures. *Mol. Ecol. Res*. 2011, 11(1): 5-18.
22. Nei M. A new measure of genetic distance (*Rapers presented at a genetics workshop during 4-th Intl. Cong. Human Genetics, Paris, 1971*). Compiled by J.F. Crow and C. Deeniston. NY: Plenum Press, 1974, 63-76.
23. Nei M. Genetic distance between populations. *Amer. Naturalist*. 1972, 106(No 949): 283-292.
24. Nei M. Analysis of gene diversity in subdivided populations. *Proc. Nat. Acad. Sci*. 1973, 70(12): 3321-3323.
25. Nei M., Roychoudhury A.K. Sampling variances of heterozygosity and genetic distance. Genetics. 1974, 76(2): 379-390.
26. Nei M. F-statistics and the analysis of gene diversity in subdivided populations. *Ann. Hum. Genet*. 977, 41(2): 225-233.
27. Nei M. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*. 1978, 89(3): 583-590.
28. Nei M., Chesser R.K. Estimation of fixation indices and gene diversities. *Ann. Hum. Genet*. 1983, 47(3): 253-259.
29. Nei M., Tajima F., TateY. Accuracy of estimated phylogenetic trees from molecular data. *J. Mol. Ed*. 1983, 19(2): 153-170.
30. Nei M. Genetic distance and molecular phylogeny. In: *Population genetics and fishery management* (N. Ryman, F. Utter, Eds). 1987, 193-223.
31. Nei M., Takezaki N. Estimation of genetic distances and phylogenetic trees from DNA analysis. *Proceedings of the World Congress on Genetics Applied to Livestock Production*, 1994, 21: 405-412.
32. Paetkau D., Waits L.P., Clarkson P.L., Craigheadg L., Strobe C. An empirical evaluation of genetic distarice statistics using microsatellite data from bear (ursidae) populations. *Genetics.* 1997, 147(4): 1943-1957.
33. Reynolds J., Weir B.S., Cockerham C.C. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics.* 1983, 105(11): 767-779.
34. Saitou N., Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 1987, 4(4). 406-425. DOI:10.1093/oxfordjournals.molbev.a040454.
35. Takezaki N., Nei M. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics*. 1996, 144(1): 389-399.
36. Weir B.S. *Analiz geneticheskikh dannykh. Metody diskretnogo analiza populyatsionno-geneticheskikh dannykh* [Genetic data analysis. Methods for discrete population genetic data]. Moscow: Mir, 1995, 400 p. (in Russian)
37. Weir B.S. Estimating F-statistics: a historical view . *Philos. Sci*. 2012, 79(5): 637-643. DOI: 10.1086 / 667904.
38. Workman P.L., Niswander J.D. Population studies on southwestern indian tribes. II. Local genetic differentiation in the Papago. *Am. J. Hum. Genet*. 1970, 22(1): 24-49.
39. Wright S. Isolation by distance. *Genetics.* 1943, 28(2): 114-138.
40. Wright S. The genetical structure of populations. *Ann. Eugenics*. 1951, 15(4): 323-354.
41. Wright S. *Evolution and the genetics of population. Vol. 4. Variability within and among natural populations*. Chicago: Univ. Chicago Press, 1978, 580 p.
42. Zhivotovsky L.A. Estimating divergence time with the use of microsatellite genetic distances: impacts of pop-ulation growth and gene flow. *Mol. Biol. Evol*. 2001, 18(5): 700-709.
43. Zhivotovsky L.A. *Populyatsionnaya biometriya* (Population biometry). Moscow: Nauka Publ., 1991, 271 p. (in Russian).
44. Zhivotovsky L.A. [Microsatellite variability in human populations and the methods of its analysis]. *VOGiS Herald.* 2006, 10(1): 74-96 (in Russian).

Nei's methods for analyzing genetic differences between populations

Kuznetsov V.M.

Rudnitsky Federal Agricultural Research Center of the North-East,
Kirov, Russian Federation

ABSTRACT. A key issue in determining and measuring population differentiation is the quantification of the nonrandom distribution of genetic variation. Studies of species divergence and genetic differentiation of populations require analysis of both heterozygosity (diversity) and genetic distances (difference), which measure different aspects of variability. Knowing how genetic variation is distributed among populations has important implications not only for evolutionary biology and ecology, but also for breeding and conservation of breeds of productive animals. There are various methods and computer programs for analyzing genetic variation by DNA markers (microsatellites, single nucleotide polymorphism), which are used to study animal populations. At the same time, the genetic and mathematical foundations of methods in Russian publications are not sufficiently reflected. Their consideration was the purpose of this work. In particular, Nei's approaches (Nei, 1974-1994) to assess genetic differences between populations based on the probability of the identity of two randomly extracted genes within and between populations are presented. In contrast to the Wright fixation index for the diallelic locus, Nei's statistics are expressed in terms of intrapopulation and interpopulation gene diversity. Formulas for calculating paired genetic distances and summary estimates of the gene differentiation of populations are presented. The numerical examples illustrate: a preliminary χ^2 test for the difference in the allelic profiles of populations, calculations of unbiased estimates of the minimum (uD_{min}), standard (uD_N) and maximum (uD_{max}) genetic distances, combined estimates of the absolute (uD_{ST}) and relative (uG_{ST}) gene differentiation, their variants and standard errors. Nei's gene diversity measures are applicable to any populations, regardless of the number of loci, the polymorphism of alleles at the locus, the presence of evolutionary factors (mutations, migration, gene drift and selection). Estimates of Nei's genetic differentiation and genetic distances by molecular genetic markers can provide valuable additional information that allows breeders, in combination with traditional and biometric methods, to make the right decisions on breeding, improving, crossbreeding and preserving breeds of productive animals.

Keywords: heterozygosity, genetic diversity, genetic distance, coefficient of gene differentiation.

Problemy biologii productivnykh zhivotnykh - Problems of Productive Animal Biology, 2020, 1: 91-110

Поступило в редакцию: 05.02.2020 Получено после доработки: 05.03.2020

Кузнецов Василий Михайлович, д.с.-х.н., проф., т.(8332)33-10-72, vm-kuznetsov@mail.ru