УДК 575+56

Искажение популяционной статистики как результат различных методических подходов к приготовлению геномных библиотек древней ДНК

Ф. С. Шарко[1*], К. В. Жур[1], В. А. Трифонов[1,2], Е. Б. Прохорчук[1]
[1]Федеральный исследовательский центр «Фундаментальные основы биотехнологии» РАН, Москва, 119071 Россия
[2]Институт истории материальной культуры РАН, Санкт-Петербург, 191186 Россия
*E-mail: fedosic@gmail.com
Поступила в редакцию 26.12.2022
Принята к печати 02.03.2023
DOI: 10.32607/actanaturae.11898

РЕФЕРАТ **На сегодняшний день существует несколько различных методов подготовки ДНК-библиотек для палеогенетических исследований. Однако химические реакции, лежащие в основе каждого из них, способны повлиять на нуклеотидную последовательность библиотек древней ДНК (дДНК) и внести изменения в результаты статистического анализа. Нами проведено сравнение результатов секвенирования библиотек дДНК образца эпохи бронзы из погребений кавказского могильника Клады. Библиотеки были приготовлены с применением трех различных подходов: метода дробовика (shotgun sequencing), отбора целевых районов генома и отбора целевых районов генома с предварительной обработкой ДНК смесью урацил-ДНК-гликозилазы (UDG) и эндонуклеазы VIII. Проанализировано влияние этих подходов на результаты вторичного статистического анализа данных, а именно F4-статистики, ADMIXTURE и метода главных компонент (PCA). Показано, что при приготовлении геномных библиотек без использования урацил-ДНК-гликозилазы возможно искажение результатов статистической обработки, связанное с посмертными химическими модификациями дДНК. Эти искажения можно нивелировать путем анализа только однонуклеотидных полиморфизмов генома, вызванных трансверсиями.**
КЛЮЧЕВЫЕ СЛОВА **древняя ДНК, ADMIXTURE, урацил-ДНК-гликозилаза, UDG.**
СПИСОК СОКРАЩЕНИЙ **UDG – урацил-ДНК-гликозилаза; дДНК – древняя ДНК; ОНП – однонуклеотидный полиморфизм; PCA – метод главных компонент.**

ВВЕДЕНИЕ

Российская Федерация является богатым источником археологического материала для проведения палеогенетических исследований. Материалы из нашей страны так или иначе использовались практически во всех громких открытиях, сделанных на основе дДНК, таких, как открытие Денисовского человека [1], популяции Ancient North Euarsians (ANE) [2], восточных охотников-собирателей (EHG) [3], популяции ямников [4]. Тем не менее сами палеогенетические исследования в России имеют весьма скромные достижения, по большей части ограничиваясь донорством костного материала [5]. Однако в последнее время создано несколько междисциплинарных коллективов, которые ставят своей целью комплексные исследования: от экспедиционных находок через синтез археологических и палеогенетических данных к генерации новых исторических гипотез. В данной работе суммированы методические подходы, опробованные в ФИЦ Биотехнологии, и выработаны наиболее эффективные алгоритмы создания геномных библиотек, которые могут быть использованы другими лабораториями, работающими в данной области.

Анализ последовательности древней ДНК (дДНК) стал мощным инструментом для изучения древних популяций человека [6–8]. Однако существует ряд сложностей, обусловленных посмертной деградацией генетического материала под действием эндогенных нуклеаз, а также случайного гидролиза и окисления. Наиболее частым повреждением дДНК является дезаминирование остатков цитозина, т.е. отщепление аминогруппы от азотистого основания с образованием остатков урацила, которые, в свою очередь, превращаются в остатки тимина в процессе проведения полимеразной цепной реакции при приготовлении библиотек фрагментов ДНК [9]. Как ре-

зультат, при секвенировании библиотек фрагментов дДНК исследователь наблюдает замены C > T на 5'-конце молекулы ДНК или G > A на 3'-конце в зависимости от особенностей выбранного протокола пробоподготовки. Наличие таких замен (ложных нуклеотидов) снижает точность картирования прочтений на референсную последовательность, при которой прочтения, содержащие нереференсные аллели, будут картированы с меньшей вероятностью, чем содержащие эталонные аллели [6].

Общее количество ложных нуклеотидов, присутствующих в реконструированном геноме, зависит от количества полученных данных секвенирования, количества накопленных посмертных повреждений ДНК и от того, проводилась ли предварительная обработка ДНК смесью ферментов урацил-ДНК-гликозилазы (UDG) и эндонуклеазы VIII (смесь позволяет вырезать урацил с образованием однонуклеотидного разрыва) при приготовлении геномных библиотек [10]. Реконструированные древние геномы обычно содержат последовательности как с подлинными, так и с искусственными вариантами, что может влиять на частотный анализ аллелей и определение структуры популяции [6].

Анализ базы данных древних геномов (Allen Ancient DNA Resource https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data) показывает, что большинство секвенированных геномов получены путем создания библиотек методом дробовика (shotgun, SG) с обработкой или без обработки UDG, либо отбором целевых районов с помощью систем обогащения. Среди последних чаще всего используют наборы производства компаний Twist, Agilent и Arbor. Системы обогащения были недавно детально охарактеризованы группой Дэвида Райха [11], что позволило выделить наборы Twist как наиболее эффективные. Однако в этой работе на затронуты такие важные вопросы, как влияние методов пробоподготовки на результаты вторичного статистического анализа. К ним мы относим F-статистики, анализ предковых компонент ADMIXTURE и многомерные проекции PCA (Principal Component Analysis).

Важную роль в эволюционной генетике человека играет изучение структуры популяций, что позволяет охарактеризовать генетическую изменчивость [12], т.е. наличие различных уровней генетического родства между некоторыми подгруппами в данной популяции. Это необходимо, когда, например, требуется сделать вывод о времени расхождения между популяциями, происходящими из разных географических мест [13, 14]. Для математического решения этой задачи делается формальное предположение о существовании так называемых «предковых популяций», из которых произошли анализируемые группы. Эти «предковые популяции» являются математической абстракцией, они характеризуются определенными частотами аллелей, и по их вкладу в анализируемые реальные образцы можно создавать компактные визуальные сводки, иллюстрирующие структуру популяции в выборке.

Алгоритмы генетической кластеризации, реализованные в таких программах, как STRUCTURE [15] и ADMIXTURE [16], широко используются для характеристики отдельных образцов и популяций на основе генетических данных. Программа ADMIXTURE эффективно оценивает индивидуальное происхождение, вычисляя оценки максимального правдоподобия в параметрической модели. Данная модель утверждает, что генотип n_{ij} для индивида i в однонуклеотидном полиморфизме (ОНП) j представляет собой количество наблюдаемых аллелей типа «1». Учитывая K предковых популяций, вероятность успеха в

$$p_{ij} = \sum_{k=1}^{K} q_{ik} f_{kj}$$

биномиальном распределении $n_{ij} \sim \mathrm{Bin}(2, p_{ij})$ зависит от доли q_{ik} происхождения i, относящейся к популяции k, и от частоты f_{kj} аллеля 1 в популяции k. ADMIXTURE максимизирует логарифмическую вероятность модели с использованием блочной релаксации:

$$L(Q,F) = \Sigma_{ij}\{n_{ij} \times \mathrm{In}p_{ij} + (2-n_{ij}) \times \mathrm{In}(1-p_{ij})\},$$

где q_{ik} и f_{kj} составляют матрицы Q и F соответственно [17].

Используемая в данной работе F4-статистика определяет вероятность независимости двух ветвей графа, где в каждой ветви находится по две популяции. В пространстве частот статистика выглядит как математическое ожидание произведения разности частот двух популяций в каждой из двух ветвей графа по всем позициям ОНП:

$$\mathrm{F4(A,B;C,D)}=\{(a-b)(c-d)\},$$

где A, B, C, D – популяции, a, b, c, d – соответствующие частоты.

Статистика считается значимой, если значение z score распределения элементов выборки F4-статистики будет по модулю больше 3. Часто в палеогенетике невозможно получить частоты популяции в силу недоступности значимого количества образцов. В этом случае аналогом F4-статистики выступает ABBA-BABA-тест по геноти-

пам единичных геномов. Для оценки вероятности отличия ABBA-BABA-теста от нуля используется также то, что величина z score должна быть по модулю больше 3, а в качестве элементов выборки берутся значения ABBA-BABA-теста для равномерных окон генома [18]. Если одна из тестируемых популяций в F4 (или в случае генотипов в ABBA-BABA-тесте) в историческом, морфологическом или генетическом аспекте очень удалена от исследуемых групп (outgroup), то тогда интуитивное объяснение ненулевой F4-статистики сводится к тому, насколько велик вклад популяции, находящейся в одной ветви с outgroup, в одну из двух популяций, находящихся в другой ветви графа.

Если абстрагироваться до представления генома образца в виде вектора, координатами которого являются значения генотипов, состоящие из трех цифр 1, 0, -1, как аналога гомозигот AA и BB и гетерозиготы, то генотипированные образцы можно представить как набор многомерных (по суммарному количеству определяемых ОНП) векторов, которые можно проецировать на пространства меньшей размерности. Один из методов проецирования – PCA, позволяет визуализировать взаимное расположение образцов. В частности, при проведении этногеографических исследований попарные расстояния между образцами современных геномов в PCA-анализе коррелируют с попарными географическими расстояниями между точками проживания доноров этого генетического материала. Нанесение древних геномов на PCA-карты, построенные в векторном пространстве современных геномов, является удобным инструментом оценки генетической связи между древними и современными людьми.

В данной работе мы сравниваем результаты популяционного ADMIXTURE и PCA-анализа, а также значения F4-статистики для образца эпохи бронзы из кавказских погребений могильника Клады (станица Царская) [19], полученные с использованием трех подходов. В качестве метода создания геномных библиотек были использованы: 1) стратегия shotgun-секвенирования; 2) отбор целевых районов генома набором компании Arbor; 3) отбор целевых районов генома набором компании Arbor с предварительной обработкой ДНК UDG; 4) отбор целевых районов генома набором компании Agilent с предварительной обработкой ДНК UDG.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Выделение ДНК и приготовление геномных библиотек

Все работы с дДНК проводили в «чистой комнате», расположенной на базе ФИЦ Биотехнологии Российской академии наук (Институт биоинженерии им. Скрябина). Использовали ДНК, полученную из антропологического образца, представляющего собой останки (фрагменты челюстей с зубами) взрослого индивида из мегалитических гробниц майкопской культуры эпохи бронзы (Северо-Западный Кавказ). ДНК выделяли из 100 мг костной пудры с использованием буфера Дабни (5 М гидрохлорид гуанидина, 40% (об./об.) 2-пропанола, 0.12 М ацетата натрия и 0.05% (об./об.) Tween 20) и магнитных шариков, покрытых силикагелем [20]. Полученную ДНК использовали для приготовления библиотек одноцепочечных фрагментов ДНК высокой сложности с помощью набора реагентов ACCEL-NGS 1S Plus DNA Library Kit (Swift Biosciences, США) в соответствии с оригинальным протоколом с некоторыми модификациями: в случае этапов ПЦР с индексацией использовали полимеразу, которая была разработана так, чтобы остатки урацила (KAPA HiFi HS Uracil+RM, США) не останавливали синтез цепи.

Из одного и того же экстракта были приготовлены три разных типа библиотек фрагментов ДНК для последующего секвенирования нового поколения. Первый (I) тип библиотеки (KLD_SG) получен с применением стратегии shotgun-секвенирования всего генома (подход, основанный на методе дробовика). Для второго (II) типа библиотеки (KLD_CAP) использовали тот же протокол подготовки с последующим обогащением по интересующим областям генома (стратегия целевого обогащения). Третий (III) тип библиотеки (KLD_UDG) отличается от типа II тем, что ДНК предварительно обрабатывали смесью урацил-ДНК-гликозилазы (UDG) и эндонуклеазы VIII, которая удаляет остатки урацила из цепей ДНК и превращает полученные абазические сайты в однонуклеотидные разрывы [21]. Обработка смесью UDG и эндонуклеазы VIII успешно удаляет остатки урацила из молекул древней ДНК, сохраняя их значительную часть на концах фрагментов, так называемые «усы древности», свидетельствующие о том, что ДНК действительно древняя. Данные по четвертому типу (IV) библиотеки, синтезированной из тех же костных останков по принципу отбора целевых районов набором компании Agilent с предварительной обработкой исходной дДНК UDG, получены от профессора Пинхасси (Университет Вены, Австрия) и обозначаются индексом I6268 из ранее опубликованной работе [22].

Целевое обогащение

Для захвата 1.6 млн ОНП из образцов древней ДНК человека мы использовали набор MyBaits Expert Human Affinities Prime Plus Kit [MyBaits Manual v.1.0 – Population Genomics Hybridization

Capture for Target NGS, 2021. https://arborbiosci.com/wp-content/uploads/ 2021/03/myBaits_Expert_HumanAffinities_v1.0_Manual.pdf]. Реагенты для обогащения по отобранным регионам генома состоят из биотинилированных одноцепочечных ДНК-зондов, которые представляют собой смесь трех наборов зондов: панель prime 1240K [23], Y Chr 46K (сайты Y-хромосомы, идентифицированные Международным обществом генетической генеалогии ISOGG) и MitoTrio (набор зондов на три различных митохондриальных генома, включая пересмотренную эталонную последовательность Кембриджа (rCRS), реконструированную эталонную последовательность Sapiens (RSRS) и последовательность неандертальца Vindija [24]). Протокол для набора MyBaits Expert Human Affinities Prime Plus предусматривает два последовательных раунда обогащения.

Секвенирование

Все три геномные библиотеки (как shotgun, так и библиотеки, обогащенные по интересующим регионам генома) были секвенированы на платформе Illumina Hiseq 4000 (1× 75 + 8 + 8 циклы) с одиночными ДНК-прочтениями.

Биоинформатическая обработка

Для фильтрации контаминирующих ДНК-прочтений из данных секвенирования мы применяли программное обеспечение BBDuk [25], входящее в пакет BBMap (www.sourceforge.net/projects/bbmap/), с использованием баз данных бактерий, грибов, растений, вирусов и «других» (http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/). Выходные данные инструмента BBDuk были обработаны с помощью конвейера PALEOMIX (версия 1.2.14) [26], в котором были обрезаны адаптеры секвенирования с помощью программы cutadapt v3.4 [27], а также выполнено сопоставление с эталонной последовательностью генома человека (hg19/GRCh37) с использованием BWA (версия 0.7.17) [28]. Выровненные прочтения отфильтровали для обеспечения минимального качества отображения 20 с помощью samtools view v1.9 [29]. Индексацию, сортировку и удаление дубликатов (rmdup) выполняли с помощью samtools v1.9 [29].

PileupCaller (https://github.com/stschiff/sequenceTools) использовали для вызова генотипов из выровненных прочтений с помощью опции «--randomHaploid», который вызывает гаплоидные генотипы путем случайного выбора одной базы высокого качества (показатель качества базы phred ⩾ 30) из панели ОНП 1240K (https://reich.hms.harvard.edu/).

Паттерны посмертных повреждений ДНК анализировали с помощью инструмента MapDamage2 [30], который предлагает несколько инструментов для визуализации и моделирования закономерностей посмертных повреждений, наблюдаемых в древних образцах. MapDamage2.0 также позволяет пересчитывать базовые показатели качества, чтобы смягчить влияние посмертного повреждения на последующие анализы.

Мы использовали программу ADMIXTURE v.1.3.0 [16] для определения генетической кластеризации образца эпохи бронзы из погребений могильника Клады (Кавказ) с помощью каждого из трех способов приготовления геномных библиотек, а также остальных образцов из панели Allen Ancient DNA Resource (AADR). ОНП были обрезаны для сайтов с неравновесным сцеплением с использованием PLINK v1.9 [31] с размером скользящего окна в 50 вариантов, размером шага – 5 вариантов и порогом r2 в 0.2 (–indep-pairwise 50 5 0.2). Мы сделали 10 повторов со случайными начальными значениями для числа кластеров (K) от 4 до 13 и выбрали прогон с минимальной ошибкой перекрестной проверки для построения графика примесей популяций.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Характеристика геномных библиотек

Вклад посмертных модификаций ДНК в искажения результатов статистического анализа мы проверяли с использованием трех геномных библиотек, приготовленных из археологического образца костей одного из захоронений могильника Клады станицы Царской: (I) shotgun-библиотека KLD_SG; (II) обогащенная целевыми районами библиотека KLD_CAP и (III) обогащенная целевыми районами и обработанная смесью урацил-ДНК-гликозилазы (UDG) и эндонуклеазы VIII, которые удаляют остатки урацила из цепей ДНК и превращают полученные абазические сайты в однонуклеотидные пробелы, библиотека KLD_UDG. Таким образом, мы ожидали, что в библиотеке I и II могут быть найдены замены C-T, которые потенциально искажают результаты генетического анализа, а в библиотеке III они искусственно удалены, но эта библиотека ожидаемо несет более короткие фрагменты, что связано с внесением одноцепочечных разрывов в исходные молекулы ДНК за счет обработки UDG.

Общее количество ДНК-прочтений, сгенерированных для этих трех ДНК-библиотек, варьировало от 58364547 до 1473546011 на ДНК-библиотеку, а количество эндогенной ДНК (т.е. прочтения, которые картировались на геном человека hg19/GRCh37) составило от 3.18 до 7.53% (*табл. 1*).

Следует отметить, что оценка количества ОНП, пригодных для анализа, проведено только по списку 1240K панели, используемой для работы с дДНК

Таблица 1. Статистика секвенирования

Библиотека	Количество входных прочтений	Количество прочтений после фильтрации	Средняя длина прочтений для анализа	Откартировалось прочтений	После удаления PCR-дубликатов	Покрытие	Эндогенная ДНК, %	ОНП (для анализа)
KLD_SG (I)	1473546011	1469259287	78.02	65025843	46813163	1.17	3.18	321229
KLD_CAP (II)	100874292	100870259	79.06	85406013	3865852	0.09	3.83	615991
KLD_UDG (III)	58364547	58329170	63.95	52565836	4392304	0.08	7.53	690148
I6268	*	*	*	*	1091304	0.81	4.02	372480

Примечание. В статье Wang et al. [22] не приведены метрики, обозначенные *.

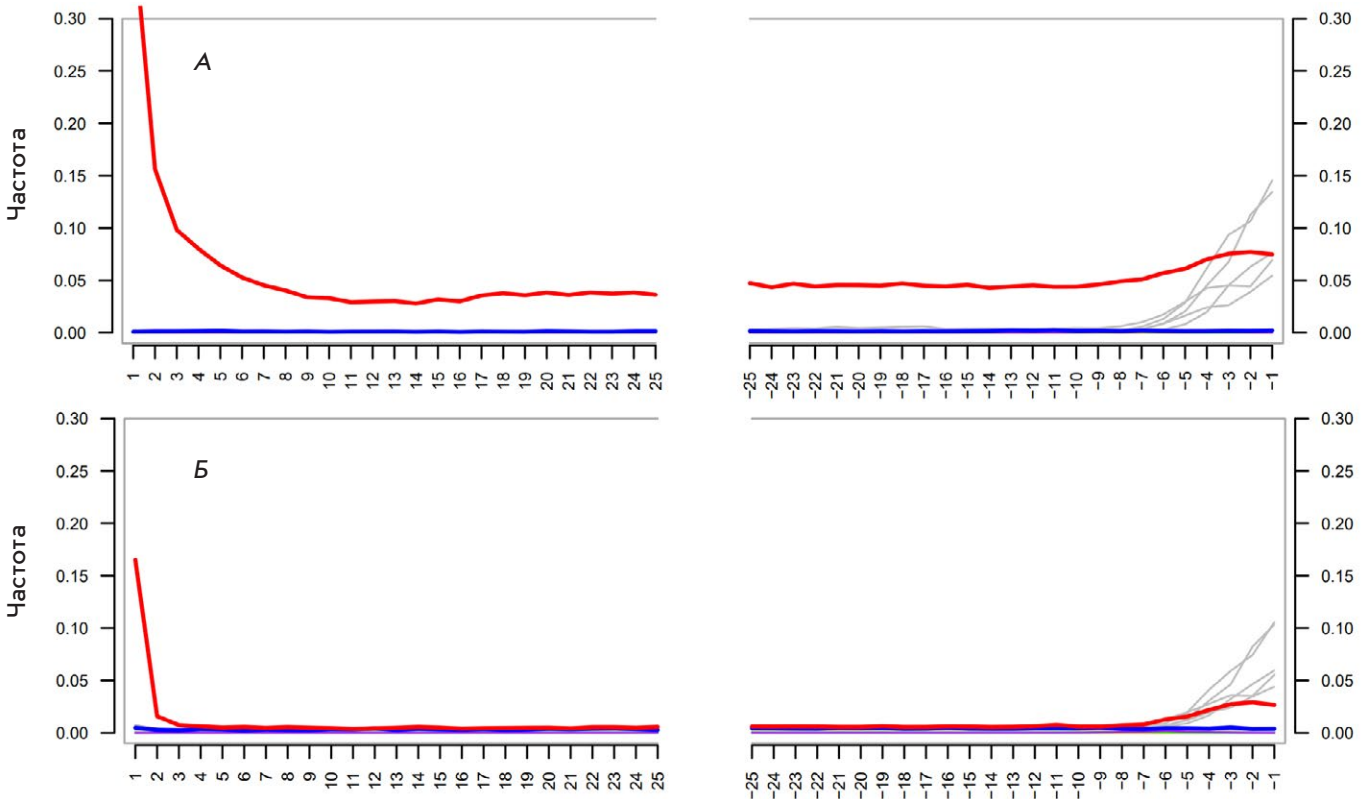

Рис. 1. Паттерны посмертных повреждений ДНК-библиотек, сгенерированные MapDamage2.0. *А* – в библиотеке, не обработанной UDG (KLD_SG), ложные переходы из C в T показаны красной линией (синей линией – комплементарные G в A) на 5′- (положительные координаты) и 3′-концах (отрицательные координаты) последних 25 нуклеотидов. Наличие замен C в T на втором прочтении (правая половина графика) комплементарной цепи обусловлено особенностями приготовления одноцепочечной библиотеки. Падение уровня дезаминирования в нуклеотидах -5, -1 связано с A-тэйлингом и также отражает особенности экспериментального протокола. *Б* – характер повреждений библиотеки (KLD_UDG), частично обработанной UDG, полученной из того же экстракта

[32]. В библиотеке I, несмотря на почти десятикратное превышение количества прочтений над библиотеками II и III, определено приблизительно в 2 раза меньше ОНП.

Подлинность древней ДНК оценивали с помощью программы MapDamage2.0, которая использует паттерны посмертных повреждений (*рис. 1*). Учитывая, что библиотеки, обработанные UDG, по-прежнему сохраняют определенное количество замен C > T на последних 2 п.н. картированных фрагментов, следует удалить по 2 п.н. с обоих концов прочтений с помощью модуля trimBam bamUtil [33].

ADMIXTURE-анализ

Результаты ADMIXTURE-анализа для K = 7 трех приготовленных в данной работе библиотек с ранее секвенированным тем же образцом новосвободненской культуры I6268 [22], к которой и принадлежат захоронения могильника Клады станицы Царской, выявили дополнительные компоненты в образцах, не обработанных UDG (KLD_CAP и KLD_SG), в виде «зеленой» и «фиолетовой» составляющей (*рис. 2А*). В силу идентичности костного образца выдвинута гипотеза о ложном влиянии посмертных модификаций дДНК на результаты ADMIXTURE при приготовлении библиотек без использования UDG.

Нами предложен биоинформатический подход к уменьшению влияния посмертных модификаций, который заключается в маскировании всех ОНП, относящихся к транзициям (C -> T и комплементарные G -> A). После удаления всех транзиций результаты ADMIXTURE-анализа библиотек, не обработанных UDG, соотносятся с результатами анализа библиотек, обработанных UDG (*рис. 2Б*).

Обращает на себя внимание тот факт, что на фоне исчезновения ложных «зеленого» и «фиолетового» компонентов при переходе к анализу трансверсий изменяется пропорциональный состав предковых популяций: увеличивается «синий» компонент, но уменьшаются «красный» и «розовый». Объяснение данного наблюдения заключается в существенном уменьшении количества ОНП, подаваемых на вход в программу ADMIXTURE. Действительно, количество трансверсий приблизительно в 5 раз меньше, чем общее количество ОНП. Детальные численные параметры ОНП для трех библиотек KLD_CAP, KLD_SG и I6268 приведены на *рис. 3*.

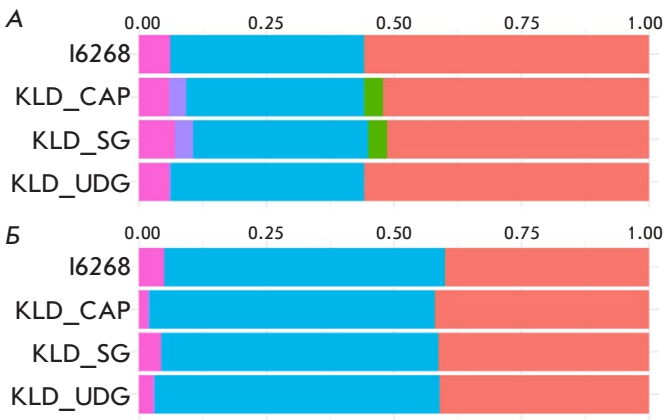

Рис. 2. ADMIXTURE-анализ (K = 7) генома, секвенированного с использованием различных методических подходов. *А* – стандартный анализ примесей, основанный на панели Дэвида Райха SNP 1240K. Можно наблюдать дополнительные базовые «зеленые» и «фиолетовые» популяции в виде дополнительных составляющих, вызванных посмертными изменениями. *Б* – анализ примесей, выполненный только на трансверсиях

Таким образом, использование трансверсий в ADMIXTURE-анализе убирает ложноположительные сигналы в виде предковых популяций, возникающих только вследствие постмортальных модификаций ДНК, а не реальных исторических популяционных перипетий. Однако необходимо с осторожностью использовать такого рода генетический редукционизм, поскольку падение общего количества входных данных за счет отсекания транзиций может повлиять на достоверность результатов анализа. Наш эмпирический опыт показывает, что порог достоверности наступает при использовании менее 30 000 ОНП.

Проанализировав все образцы из панели Allen Ancient DNA Resource (AADR) V44.3 (январь 2021), мы выявили значительную отрицательную корреляцию (-0.5844) между дополнительным компонентом ADMIXTURE и наличием обработки UDG в 3284 европейских образцах из базы данных Allen Ancient DNA Resource (AADR) (*рис. 4*). Следует также отметить, что использование процедуры обогащения целевыми районами приводит к экономии ресурсов при секвенировании. Действительно, генерация 58 млн прочтений в библиотеке KLD_UDG приводит к выводам о структуре предковых популяций, сравнимым с выводами, сделанными с использованием библиотеки KLD_SG (около 1500 млн прочтений).

F4-статистика

Для исследования роли пробоподготовки в интерпретации результатов популяционного анализа нами была рассчитана F4-статистика в конфигурации F4(Wang_3, Y;X, Yoruba). Популяция Wang_3 состоит из трех образцов (I6267, I6266 и I6272) новосвободненской культуры, к которой и принадлежит также образец I6268. Популяция X взята из списка, предложенного археологами (приведен на *рис. 5* слева от оси ординат), в контексте их исторических гипотез. Популяцию Yoruba использовали в качестве outgroup. В качестве популяций Y выступали четыре набора ОНП, определенных в KLD_SG, KLD_CAP, KLD_UDG и I6268. Как сказано выше, интуитивный смысл ненулевой достоверной статистики укажет на ту популяцию из списка X, которая вносит больший вклад в популяцию Wang_3 в случае положительной статистики и больший вклад в экспериментальный образец Y в случае отрицательной статистики. Интерпретация археологического и исторического смысла разницы между популяцией Wang_3 и образцом, использованным для приготовления четырех тестовых библиотек,

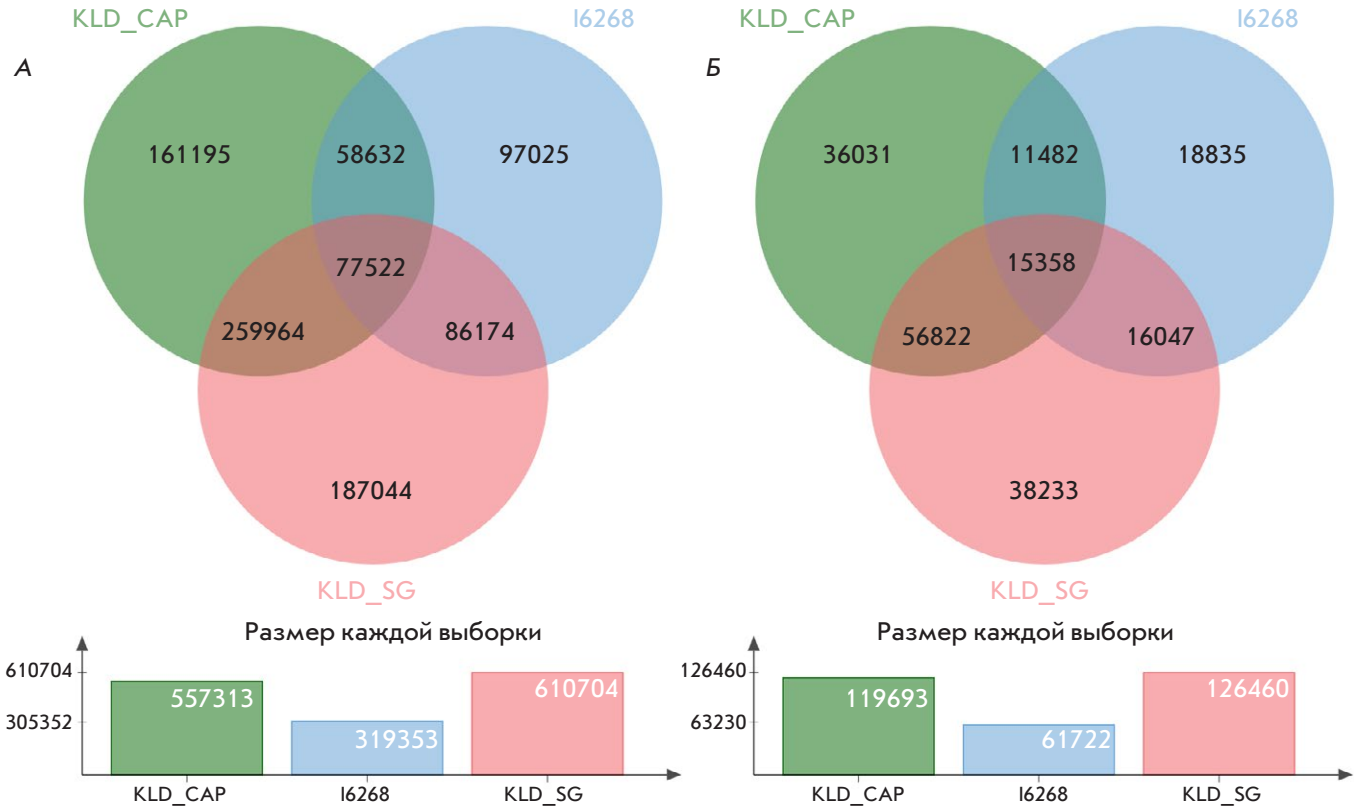

Рис. 3. Диаграммы Венна для ОНП трех библиотек KLD_CAP, KLD_SG и I6268. Общее количество ОНП для каждой библиотеки приведено под диаграммами. *А* – трансверсии и транзиции; *Б* – трансверсии

выходит за рамки данной статьи. Расчеты приведены лишь для примера, чтобы указать на возможности различной интерпретации достоверной ненулевой F4-статистики в зависимости от метода пробоподготовки.

На *рис. 5А* приведены данные F4-статистики для всех ОНП. При использовании в качестве Y KLD_SG все популяции справа отсортированы по убыванию значения статистики. При использовании в качестве Y трех других библиотек порядок сортировки существенно меняется. Более того, достоверные статистики с модулем z score больше 3 изменяются: 12, 9, 8 и 2 для KLD_SG, KLD_CAP, KLD_UDG и I6268 популяций соответственно. Всего лишь одна популяция из списка X – Russia_HG_Tyumen – является достоверной для всех четырех библиотек списка Y. Однако при переходе к анализу трансверсий порядок сортировки всех четырех библиотек меняется по сравнению с исходной сортировкой F4 (Wang_3; KLD_SG; X; Yoruba). Количество достоверных популяций из списка X составляет 7, 9, 4 и 1 для KLD_SG, KLD_CAP, KLD_UDG и I6268 соответственно. Но при этом нет ни одной популяции X, которая бы достоверно определялась через F4-статистику во всех четырех библиотеках при ис-

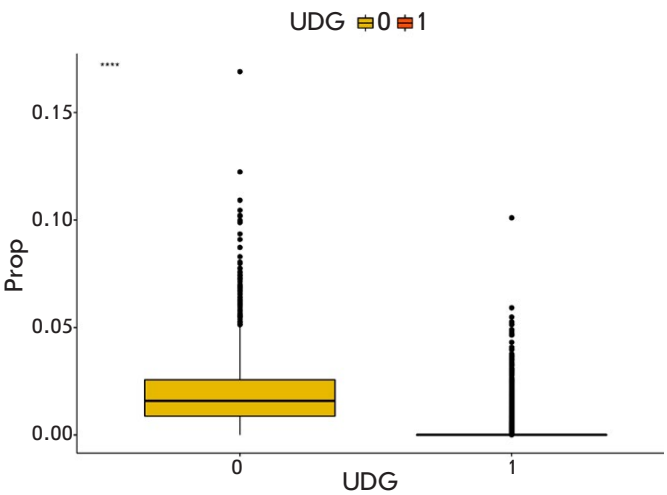

Рис. 4. Распределение относительного содержания дополнительных составляющих в ADMIXTURE-анализе в образцах из панели Allen Ancient DNA Resource V44.3, обработанных UDG (1) – 2376 образцов и без обработки (0) – 908 образцов. Prop – пропорция дополнительных примесей

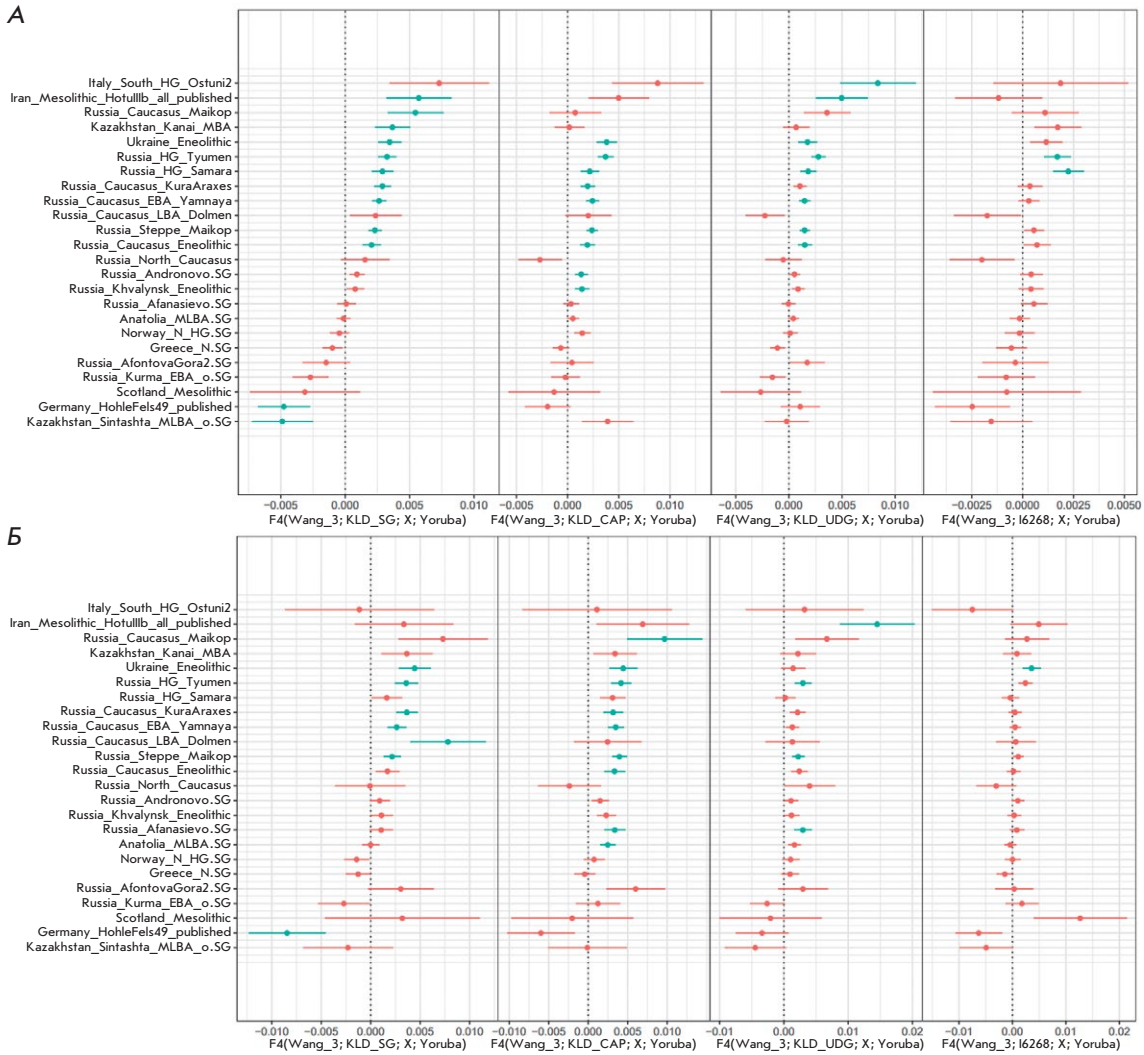

Рис. 5. F4-статистика в конфигурации F(Wang_3, Y;X, Yoruba). Синим обозначены достоверные метрики при $|z| > 3$. *А* – F4, рассчитанная на всех ОНП из панели 1240K. *Б* – F4 только на трансверсиях

пользовании трансверсий. Как видно из *рис. 5Б*, даже Russia_HG_Tyumen, будучи общей достоверной популяцией X при анализе всех ОНП, является достоверной лишь в трех библиотеках – KLD_SG, KLD_CAP, KLD_UDG, но не в I6268.

Вывод из этой части работы заключается в том, что F4-статистика по аналогии с ADMIXTURE чрезвычайно чувствительна к количеству поданных на вход ОНП, также критически важно использовать в F4-статистике наборы ОНП, полученные из единообразно приготовленных геномных библиотек. В ином случае есть вероятность неправильной интерпретации достоверных и положительных по модулю значений F4.

PCA-генетические карты

Оценено также влияние пробоподготовки на PCA-проекции на плоскость PC1-PC2. В PCA, исходно построенном для векторов представителей различных современных популяций Евразии, использованы 253 древних образца [34]. С целью упрощения восприятия все древние образцы на *рис. 6* были окрашены в светло-серый цвет за исключением четырех исследуемых библиотек. Образцы I6268 и KLD_UDG имеют минимальную разницу в координатах PC1-PC2, в то время как KLD_SG несколько отдаляется от них в направлении «северо-восток». Чтобы получить детальное представление о том, как группируются четыре тестовые библиотеки при анализе всех ОНП и только трансверсий, проведен PCA-анализ с использованием только 17 образцов исторического контекста новосвободненской культуры (*рис. 7*). Следует пояснить, что новый PCA с 17 образцами предусматривает генерацию новых векторов PC1-PC2, отличных от полученных на *рис. 6*. Видно, что ни в случае использования всех ОНП (*рис. 7А*), ни в случае использования только трансверсий (*рис. 7Б*) не удается свести хотя бы две любые библиотеки в одну точку на плоскости PC1-PC2. Надо признать, что Fst для группы KLD_SG, KLD_CAP,

Рис. 6. Метод главных компонент (PCA) с помощью проецирования «древних» образцов на вектора современных образцов, которые использовались только для построения PCA и не представлены на данном рисунке. Сначала рассчитываются главные компоненты в векторах современных образцов, после чего на них проецируются вектора «древних» образцов. На рисунке представлены только «древние» образцы в координатах PC1 и PC2 (первые две главные компоненты)

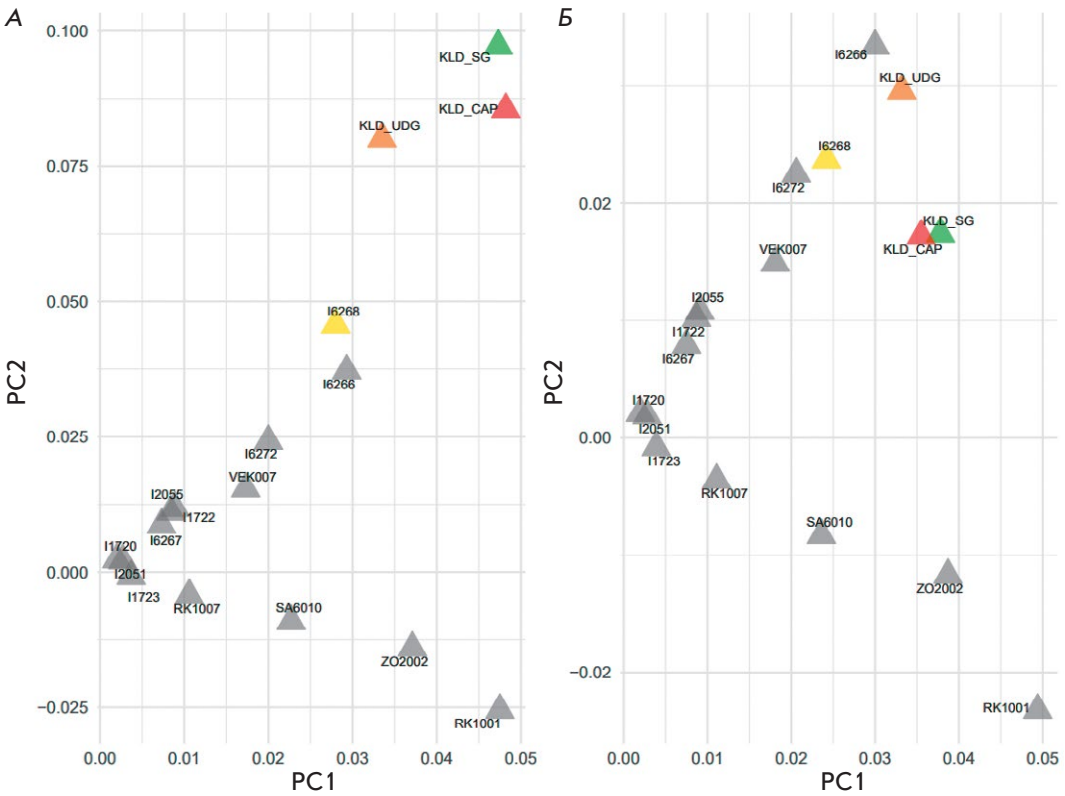

Рис. 7. Метод главных компонент (PCA) для KLD_SG, KLD_CAP, KLD_UDG, I6268 (выделены цветом) и других кавказских образцов (серые), рассчитанные на всех ОНП из панели 1240K (*А*) и только на трансверсиях (*Б*)

KLD_UDG и I6268 составляет 11% при использовании всех ОНП и 19% в случае только трансверсий, говоря о повышении конкордантности образцов при использовании трансверсий, что нашло отражение в несколько большей кучности четырех тестовых библиотек на PCA-картах.

ВЫВОДЫ

В работе показано, что современные статистические методы, особенно F4-статистика, весьма чувствительны к используемому методу пробоподготовки геномных библиотек дДНК. Оптимальным подходом к созданию геномных библиотек мы посчитали отбор целевых районов с предварительной обработкой UDG исходной дДНК. Даже при таком подходе использование наборов для обогащения целевыми районами от различных производителей может генерировать ложноположительные результаты при статистическом анализе. Для уменьшения влияния методических подходов рекомендуется увеличивать экспедиционные выборки костных останков представителей одной культуры/популяции, а также унифицировать по возможности пробоподготовку в рамках одного исследования. ●

Работа выполнена при финансовой поддержке проекта Минобрнауки России, системный номер № 075-10-2020-116 (номер гранта 13.1902.21.0023).

СПИСОК ЛИТЕРАТУРЫ

1. Krause J., Fu Q., Good J.M., Viola B., Shunkov M.V., Derevianko A.P., Pääbo S. // Nature. 2010. V. 464. № 7290. P. 894–897.
2. Long J. // Hum. Biol. 2017. V. 89. № 4. P. 303–304.
3. Haak W., Lazaridis I., Patterson N., Rohland N., Mallick S., Llamas B., Brandt G., Nordenfelt S., Harney E., Stewardson K., et al. // Nature. 2015. V. 522. № 7555. P. 207–211.
4. Morgunova N. // Radiocarbon. 2013. V. 55. № 3–4. P. 1286–1296.
5. Grigorenko A.P., Borinskaya S.A., Yankovsky N.K., Rogaev E.I. // Acta Naturae. 2009. V. 1. № 3. P. 58–69.
6. Orlando L., Allaby R., Skoglund P., Der Sarkissian C., Stockhammer P.W., Ávila-Arcos M.C., Fu Q., Krause J., Willerslev E., Stone A.C., et al. // Nat. Rev. Methods Primers. 2021. V. 1. № 1. P. 14.
7. Sokolov A.S., Nedoluzhko A.V., Boulygina E.S., Tsygankova S.V., Sharko F.S., Gruzdeva N.M., Shishlov A.V., Kolpakova A.V., Rezepkin A.D., Skryabin K.G., et al. // J. Archaeol. Sci. 2016. V. 73. P. 138–144.
8. Erlikh V.R., Gak E.I., Kleshchenko A.A., Sharko F.S., Boulygina E.S., Tsygankova S.V., Slobodova N.V., Rastorguev S.M., Nedoluzhko A., Godizov G.L., et al. // J. Archaeol. Sci. Repts. 2021. V. 39. P. 103198.
9. Axelsson E., Willerslev E., Gilbert M.T.P., Nielsen R. // Mol. Biol. Evol. 2008. V. 25. № 10. P. 2181–2187.
10. Llamas B., Valverde G., Fehren-Schmitz L., Weyrich L.S., Cooper A., Haak W. // STAR: Sci. Technol. Archaeol. Res. 2017. V. 3. № 1. P. 1–14.
11. Rohland N., Mallick S., Mah M., Maier R., Patterson N., Reich D. // Genome Res. 2022. V. 32. № 11–12. P. 2068–2078.
12. Hellenthal G., Busby G.B.J., Band G., Wilson J.F., Capelli C., Falush D., Myers S. // Science. 2014. V. 343. № 6172. P. 747–751.
13. Gopalan S., Smith S.P., Korunes K., Hamid I., Ramachandran S., Goldberg A. // Philos. Trans. R. Soc. Lond. B Biol. Sci. 2022. V. 377. № 1852. P. 20200410.
14. Sjödin P., McKenna J., Jakobsson M. // Genetics. 2021. V. 217. № 4. iyab008.
15. Hubisz M.J., Falush D., Stephens M., Pritchard J.K. // Mol. Ecol. Resour. 2009. V. 9. № 5. P. 1322–1332.
16. Alexander D.H., Novembre J., Lange K. // Genome Res. 2009. V. 19. № 9. P. 1655–1664.
17. Zhou H., Alexander D., Lange K. // Stat. Comput. 2011. V. 21. № 2. P. 261–273.
18. Sinharay S. International encyclopedia of education. Amsterdam: Elsevier, 2010. P. 229–231.
19. Шишлина Н.И., Трифонов В.А., Хоммель П. // Краткие сообщения Института археологии (КСИА). 2019. № 257. С. 35–47.
20. Rohland N., Glocke I., Aximu-Petri A., Meyer M. // Nat. Protoc. 2018. V. 13. № 11. P. 2447–2461.
21. Gansauge M.-T., Meyer M. // Nat. Protoc. 2013. V. 8. № 4. P. 737–748.
22. Wang C.-C., Reinhold S., Kalmykov A., Wissgott A., Brandt G., Jeong C., Cheronet O., Ferry M., Harney E., Keating D., et al. // Nat. Commun. 2019. V. 10. № 1. P. 590.
23. Mathieson I., Lazaridis I., Rohland N., Mallick S., Patterson N., Roodenberg S.A., Harney E., Stewardson K., Fernandes D., Novak M., et al. // Nature. 2015. V. 528. № 7583. P. 499–503.
24. Green R.E., Malaspinas A.-S., Krause J., Briggs A.W., Johnson P.L.F., Uhler C., Meyer M., Good J.M., Maricic T., Stenzel U., et al. // Cell. 2008. V. 134. № 3. P. 416–426.
25. Bushnell B., Rood J., Singer E. // PLoS One. 2017. V. 12. № 10. P. e0185056.
26. Schubert M., Ermini L., Der Sarkissian C., Jónsson H., Ginolhac A., Schaefer R., Martin M.D., Fernández R., Kircher M., McCue M., et al. // Nat. Protoc. 2014. V. 9. № 5. P. 1056–1082.
27. Martin M. // EMBnet J. 2011. V. 17. № 1. P. 10.
28. Li H., Durbin R. // Bioinformatics. 2009. V. 25. № 14. P. 1754–1760.
29. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. // Bioinformatics. 2009. V. 25. № 16. P. 2078–2079.
30. Jónsson H., Ginolhac A., Schubert M., Johnson P.L.F., Orlando L. // Bioinformatics. 2013. V. 29. № 13. P. 1682–1684.
31. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., Maller J., Sklar P., de Bakker P.I.W., Daly M.J., et al. // Am. J. Hum. Genet. 2007. V. 81. № 3. P. 559–575.
32. Allentoft M.E., Sikora M., Sjögren K.-G., Rasmussen S., Rasmussen M., Stenderup J., Damgaard P.B., Schroeder H., Ahlström T., Vinner L., et al. // Nature. 2015. V. 522. № 7555. P. 167–172.
33. Jun G., Wing M.K., Abecasis G.R., Kang H.M. // Genome Res. 2015. V. 25. № 6. P. 918–925.
34. Triska P., Chekanov N., Stepanov V., Khusnutdinova E.K., Kumar G.P.A., Akhmetova V., Babalyan K., Boulygina E., Kharkov V., Gubina M., et al. // BMC Genet. 2017. V. 18. № Suppl. 1. P. 110.