ГЕНЕТИКА, 2023, том 59, № 7, с. 828–838

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ

УДК 57.081.23

ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА ВОЗМОЖНОСТИ ВЫЯВЛЕНИЯ КРОСС-КОНТАМИНИРОВАННЫХ ОБРАЗЦОВ ДНК НА ОСНОВЕ ГЕНЕТИЧЕСКИХ ДАННЫХ[1]

© 2023 г. Н. В. Фелиз[1, *, #], К. С. Грамматикати[1, #], С. И. Митрофанов[1, #], П. А. Гребнев[1], К. Д. Конуреева[1], Е. Д. Маралова[1], М. В. Ерохина[1], Т. А. Шпакова[1], П. Г. Казакова[1], Ю. Н. Ахмерова[1], А. А. Мкртчян[1], Е. А. Снигирь[1], В. С. Юдин[1], А. А. Кескинов[1], С. М. Юдин[1], В. И. Скворцова[2]

[1]*Федеральное государственное бюджетное учреждение "Центр стратегического планирования и управления медико-биологическими рисками здоровью" Федерального медико-биологического агентства, Москва, 119121 Россия*

[2]*Федеральное медико-биологическое агентство, Москва, 123182 Россия*

**e-mail: feliz08nv@gmail.com*

Поступила в редакцию 15.12.2022 г.
После доработки 30.01.2023 г.
Принята к публикации 01.02.2023 г.

Проблемы кросс-контаминации и неправильной маркировки образцов биоматериала являются крайне актуальными при проведении массовых генетических исследований. В настоящем исследовании проведена экспериментальная оценка возможности выявления кросс-контаминированных образцов ДНК с использованием нескольких подходов: расчета отношения ридов, приходящихся на референсный или альтернативный аллель (allele ratio, AR); отношения количества гетерозиготных вариантов к гомозиготным; значения показателя CallRate для данных, полученных с помощью ДНК-микрочипов; программы Picard CrosscheckFingerprints (CrossCheck). Для проведения исследований были созданы контаминированные образцы (смеси) путем смешивания обычных "чистых" образцов ДНК в разных соотношениях. Показатели качества образцов проанализированы по данным полногеномного секвенирования и генотипирования с помощью ДНК-микрочипа Illumina microarray BeadArray technology CoreExome (CE). Экспериментально установлено, что все указанные подходы могут быть использованы для выявления ошибок генотипирования, связанных с контаминированием образцов.

Ключевые слова: полногеномное секвенирование, контаминация, ДНК-микрочипы, контроль качества.

DOI: 10.31857/S0016675823060061, **EDN:** SSDAUO

Для получения точных и корректных результатов генетических исследований необходимо быть уверенным в высоком качестве секвенирования, в отсутствии неправильно маркированных образцов и контаминации. Использование алгоритмов, позволяющих выявлять образцы биоматериала, поступившие от одного и того же донора, а также оценивать степень загрязненности другими биообразцами, актуально и полезно в рутинной практике.

В то время как программы для оценки качества данных секвенирования на этапе прочтений (ридов, FASTQ-файлов) и выравнивания (BAM-файлов) хорошо известны и широко применяются, распространенного метода для оценки качества итоговых данных (vcf-файлов) нет. В настоящей работе рассматриваются два подхода: расчет соотношения прочтений, приходящихся на референсный или альтернативный аллель, к общему числу прочтений (allele ratio, AR) и расчет соотношения количества гетерозиготных вариантов к гомозиготным (Het/Hom) [1]. Ранее было показано, что показатель Het/Hom различается в разных этносах, но постоянный для любых регионов генома [2]. В своей работе J. Wang и коллеги показали, что для европейцев отношение Het/Hom в среднем составляет 1.6, самое высокое значение Het/Hom наблюдается у африканцев (около 2.0), а самое низкое (около 1.4) свойственно азиатам.

[1] Дополнительная информация для этой статьи доступна по doi 10.31857/S0016675823060061 для авторизованных пользователей.

[#] Вклад этих авторов в работу равнозначный.

Одной из программ, позволяющей выявить образцы биоматериала, поступившие от одного и того же донора, является программа Picard Crosscheck Fingerprints (CrossCheck) от The Broad Institute ("Picard Toolkit" 2019, GitHub Repository https://broadinstitute.github.io/picard/). Отличительной особенностью этой программы является выдача результата в виде количественной характеристики – степени сходства образцов. При этом было показано, что CrossCheck точен и универсален для различных типов данных, таких как секвенирование РНК, ДНК и ChiP-seq [3]. Однако информация о корректности работы CrossCheck в случае контаминации отсутствует. В рамках настоящей работы мы оценили способность программы CrossCheck выявлять пары образцов биоматериала, поступивших от одного донора, при условии контаминации одного или обоих образцов.

Принцип работы программного обеспечения (ПО) CrossCheck состоит в следующем: из VCF-файлов отбираются генетические варианты и создается "генетический отпечаток". Специально для CrossCheck был создан набор из 58894 однонуклеотидных генетических вариантов (SNPs) (https://github.com/naumanjaved/fingerprint_maps/blob/master/map_files/hg38_chr.map). Эти варианты отобраны таким образом, чтобы частота встречаемости минорного аллеля (MAF) была больше 10% по данным проекта "1000 геномов". В сформированный набор вошли только биаллельные SNPs. Кроме того, все отобранные SNPs находились в блоках неравновесного наследования (Linkage disequilibrium, LD) с высоким коэффициентом корреляции внутри блока ($r^2 > 0.85$) и низким процентом корреляции между блоками ($r^2 < 0.1$). Также в разных популяциях встречаемость этих SNPs различалась менее чем на 10% по данным проекта "1000 геномов".

При сравнении "генетических отпечатков" между собой ПО CrossCheck рассчитывает метрику LOD (LOD Score) для каждой пары образцов, показывающую какие образцы с какой вероятностью имеют одинаковое биологическое происхождение. Авторами программы CrossCheck указаны следующие пороговые значения LOD: разные образцы (LOD < –5), дубли (LOD > 5), "сомнительные" (–5 < LOD < 5).

В рамках данного исследования была смоделирована контаминация выделенной ДНК и проанализированы показатели качества контаминированных образцов на данных WGS и при генотипировании с помощью технологии Illumina microarray BeadArray technology CoreExome (CE). В работе T. Dallavilla показано, что с помощью расчета AR можно выявить контаминацию свыше 10%, поэтому для данной работы нами были выбраны следующие значения контаминации: 10, 20, 30 и 50% [1]. Помимо того, что контаминацию ниже 10% нельзя обнаружить с помощью расчета AR, следует отметить, что низкая степень контаминации окажет меньшее влияние на результаты секвенирования, нежели контаминация свыше 10%.

МАТЕРИАЛЫ И МЕТОДЫ

Сбор биоматериала и метаданных

В настоящем исследовании использованы образцы цельной крови от семи доноров из коллекции ФГБУ "ЦСП" ФМБА России. Для всех биообразцов соблюдены следующие условия: наличие и корректность метаданных каждого донора (пол, возраст, регион проживания, национальность, анамнез); обеспечение правильного забора, транспортировки и хранения биоматериала (венозная кровь) в соответствии с ГОСТ Р53079.4-2008. Все биообразцы прошли проверку на отсутствие признаков гемолиза и хилеза.

Выделение ДНК из цельной крови

Выделение геномной ДНК из образцов цельной крови проводилось с использованием автоматизированной станции Tecan Freedom EVO (Tecan, Швейцария) при помощи набора MagAttract HMW DNA Kit (Qiagen, Германия), а также применялась процедура ручного выделения с применением набора DNA Blood Mini Kit (Qiagen, Германия) в соответствии с протоколом производителя.

Концентрацию и чистоту выделенной ДНК определяли двумя методами: в автоматическом режиме на планшетном ридере Infinite 200 Pro (Tecan, Швейцария) с помощью программы Magellan, а также в ручном режиме с помощью флуориметра Qubit 4.0 (Thermo Fisher Scientific, США), чистоту препаратов ДНК оценивали с помощью NanoDrop One C Microvolume UV-Vis (Thermo Fisher Scientific).

Для всех образцов выделенной ДНК отношение показателей поглощения 260/280 составляло 1.8–2.0, отношение 260/230 – 2.0–2.2.

Подготовка геномных библиотек и секвенирование образцов ДНК

Для приготовления библиотек использовали 150–500 нг геномной ДНК. Тагментацию ДНК, очистку и амплификацию тагментированнной ДНК, очистку полученных библиотек проводили согласно протоколу Nextera DNA Flex Library Prep (Document #1000000025416 v07, Illumina, США). Полногеномные библиотеки готовили с использованием набора реагентов Nextera DNA Flex kit (Illumina, США) согласно рекомендациям производителя и набора индексов IDT-ILMN Nextera DNA UD Indexes Set A и Set B для предотвращения кросс-контаминации образцов. Концентра-

Таблица 1. Названия смесей, участвующих в эксперименте

Соотношение	Образцы			
	50022	50054	50027	50005
	50016	50027	50053	50008
10 : 90	50022_10_50016_90	50054_10_50027_90	50027_10_50053_90	50005_10_50008_90
20 : 80	50022_20_50016_80	50054_20_50027_80	50027_20_50053_80	50005_20_50008_80
30 : 70	50022_30_50016_70	50054_30_50027_70	50027_30_50053_70	50005_30_50008_70
50 : 50	50022_50_50016_50	50054_50_50027_50	50027_50_50053_50	50005_50_50008_50

ция библиотек измерялась на планшетном ридере Infinite F Nano Plus (Tecan, Швейцария). Размер полученных библиотек определяли при помощи набора реагентов Agilent D1000 на приборе Agilent 4200 TapeStation (Agilent Technologies, США). Пулирование производилось автоматически с использованием роботизированной станции Tecan Freedom EVO (Tecan). Пул библиотек перед секвенированием разводили до концентрации 1.5–2.0 нМ. Контроль качества пула проводили при помощи набора реагентов Agilent HS D1000 ScreenTape на приборе Agilent 4200 TapeStation (Agilent Technologies, США). Полногеномное секвенирование проводили на приборе Illumina NovaSeq 6000 (Illumina), используя набор реагентов S4, 300 циклов (Illumina) с парно-концевыми прочтениями 2 × 150 пн.

Генотипирование с помощью технологии Illumina microarray (BeadArray technology)

Микрочиповое генотипирование проводили с использованием набора Illumina Infinium CoreExome-24 v1.3 (Illumina) по протоколу производителя (документ Infinium HTS Assay Reference Guide (15045738 v04)). Пробоподготовка проводилась с помощью автоматизированной станции Tecan Freedom EVO. Микрочипы сканировали на системе Illumina iScan с модулем автоматической подачи Autoloader 2.x.

Создание контаминированных образцов

Для проведения исследования из семи образцов выделенной ДНК было сформировано четыре пары так, что один образец (50027) присутствовал в двух парах. Для каждой пары было приготовлено четыре смеси в соотношениях 10 : 90, 20 : 80, 30 : 70 и 50 : 50. В табл. 1 представлены названия получившихся смесей и соотношения, в которых смешивались исходные образцы.

Биоинформатическая обработка WGS

Первым этапом обработки сырых данных секвенирования является процесс демультиплексации, при котором исходная выдача секвенатора NovaSeq 6000 из формата BCL конвертируется в формат FASTQ с помощью программного обеспечения bcl2fastq v2.20. Для проведения контроля качества секвенирования всей ячейки в целом использовалась программа Illumina Sequencing Analysis Viewer v2.4.7. Следующий этап биоинформатической обработки подразумевает выравнивание на референсный геном, что осуществлялось с помощью DRAGEN [4]. В качестве референсного генома использовалась последовательность GRCh38.d1.vd1 (GDC Reference Files, NCI Genomic Data Commons, https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files (accessed: 21.02.2022)).

Малые генетические варианты (SNPs) определяли с помощью программы Strelka v2.9.10 [5] с фильтром “PASS”. Количество гомо- и гетерозигот рассчитывали с помощью программы bcftools v1.14 [6].

Биоинформатическая обработка CE

Алгоритм обработки данных CE и получения VCF-файлов из результатов сканирования Illumina microarray CoreExome сводился к следующему: для конвертации файлов формата idat в формат gtc использовали Illumina Array Analysis Platform gencall v1.1 с манифест-файлами, предоставленными компанией Illumina (v1.3); для конвертации файлов формата gtc в файлы формата VCF использовали программу GTCtoVCF v1.2.1 (Illumina) с референсным геномом человека human_g1k_v37; для перевода координат SNPs с референса human_g1k_v37 на референс GRCh38.d1.vd1 использовали программу CrossMap v0.5.4 [7]. Из полученных VCF-файлов удаляли все мутации со значением 0 в поле GQ или с генотипом “.” и “./.”.

Расчет метрик для гетерозиготных вариантов

Расчет соотношения прочтений, приходящихся на референсный или альтернативный аллель, к общему числу прочтений (allele ratio, AR) и всех связанных с этим показателем метрик, а также фильтрацию мутаций проводили по методике, описанной в работе T. Dallavilla [1]. Из данных

WGS были отобраны только биаллельные гетерозиготные варианты с качеством выравнивания (map quality, MQ) более 18 и суммарным числом прочтений более 10. Для поиска AR для каждого гетерозиготного варианта рассчитали отношение количества прочтений с альтернативным аллелем к общему количеству прочтений в данной позиции.

Для 500 образцов WGS с покрытием более 30x и отношением Het/Hom в диапазоне 1.64–1.7 рассчитали стандартное отклонение AR (0.09). На основании этих данных определили "референсный" 95%-ный доверительный интервал (95%ДИ) для AR, равный $0.5 \pm 1.96 \times 0.09$.

Затем для каждого образца рассчитали процент гетерозигот с AR за пределами диапазона 0.32–0.68 и стандартное отклонение AR.

Запуск CrossCheck

Для VCF-файлов, полученных на данных WGS и CE, попарно во всех возможных комбинациях определили значение LOD Score с помощью программы PICARD CrosscheckFingerprints (v2.26.11). В качестве набора гаплотипов использовался файл, описанный в работе N. Javed [3].

РЕЗУЛЬТАТЫ

Для дальнейшего изложения введены следующие обозначения: чистый образец – образец, в котором присутствует ДНК только от одного человека; смесь образцов – образец, состоящий из биоматериала двух человек; мажорный компонент смеси – ДНК, содержание которой превышает 50% в смеси из двух образцов; минорный компонент смеси – ДНК, содержание которой менее 50% в смеси из двух образцов. В зависимости от содержания минорного компонента будем использовать термины: 10-, 20-, 30-, 50%-ная смесь. Процентное содержание минорного компонента в смеси будем называть концентрацией. Все образцы (и "чистые", и смеси образцов) генотипировали с помощью двух технологий: WGS и CE.

Показатели качества контаминированных образцов

Для всех полученных файлов сравнили метрики качества. С помощью программы CrossCheck попарно сравнили VCF-файлы, полученные с применением только одной технологии (WGS–WGS, CE–CE) и с помощью двух разных технологий (WGS–CE). Выявление малых герминальных вариантов в данных полногеномного секвенирования образца состоит из нескольких этапов. На первом этапе данные, поступающие из секвенатора в формате BCL, конвертируются в формат FASTQ. В этом формате содержится информация как о нуклеотидной последовательности каждого рида, так и о качестве, с которым определен каждый нуклеотид в прочтении. Мы оценили качество всех образцов (смесей и чистых) с помощью программы fastqc (FASTQC. A quality control tool for high throughput sequence data, BibSonomy [Electronic resource]. URL: https://www.bibsonomy.org/bibtex/f230a919c34360709aa298734d63dca3 (accessed: 31.01.2022)). Как и ожидалось, в формате FASTQ показатели качества у смесей и чистых образцов не отличаются друг от друга (Приложение).

На втором этапе обработки данных осуществляется выравнивание прочтений на референсный геном. Выравнивание проводили при помощи программы DRAGEN. Сравнение метрик качества выравнивания у смесей и чистых образцов также не выявило различий (Приложение, табл. 1 и 2).

Для выявления малых герминальных генетических вариантов использовалась программа Strelka. Одним из показателей качества является отношение количества гетерозиготных вариантов к гомозиготным (Het/Hom).

У чистых образцов на данных WGS рассчитанное отношение Het/Hom находилось в интервале 1.6–1.8, что согласуется с ранее опубликованными результатами [2], поскольку все образцы в настоящем исследовании были получены от европейцев. Для 10%-ной смеси отношение Het/Hom равно 1.9, для 20% – превышает 2.1, для 30% – выше 3.0, для 50% – выше 3.5 (табл. 2). В случае с CE наблюдалась похожая ситуация: отношение Het/Hom у 10%-ных смесей незначительно выше, чем у чистых образцов, а у смесей с содержанием минорного компонента 20% и более отношение Het/Hom превышает 2.0. Так как значение Het/Hom в диапазоне 2–2.1 является нормой для африканцев, то без информации об этнической принадлежности доноров биоматериала считать образцы с Het/Hom > 2.0 контаминированными нельзя.

Еще одна метрика, которую можно использовать для определения контаминации, – это отношение прочтений, которые приходятся на референсный или альтернативный аллель, к общему числу прочтений в конкретной гетерозиготной мутации (allele ratio, AR). В данном исследовании в чистых образцах гетерозиготные варианты с AR вне 95%ДИ не превышает 10%, тогда как у 10%-ных смесей превышают 16%, для 20%-ных – 35%, для 30%-ных – 49.5%, а для 50%-ных – 61% (табл. 2). С увеличением количества примеси увеличивается и стандартное отклонение, принимающее значение 0.1 для чистых образцов и 0.2 для 50%-ных смесей; кроме того, меняется характер распределения AR (Приложение, рис. 1).

Для оценки качества результатов генотипирования с помощью CE производителем рекомен-

Таблица 2. Метрики Call Rate и AR, отношение количества гетерозиготных вариантов к гомозиготным в образцах

Образец	Het/Hom WGS	Het/Hom CE	Call Rate CE	Стандартное отклонение (SD)	Кол-во гетерозиготных вариантов	Кол-во гетерозиготных вариантов с AR, входящих в 95%ДИ	Процент гетерозиготных вариантов с AR, не входящих в 95%ДИ
50005	1.7	1.7	1.00	0.09	2203286	2032090	7.77
50008	1.7	1.6	1.00	0.09	2194958	2030601	7.49
50016	1.7	1.7	1.00	0.09	2311041	2137817	7.50
50022	1.6	1.7	0.99	0.07	2278578	2197124	3.57
50027	1.7	1.6	1.00	0.09	2215825	2076111	6.31
50053	1.7	1.6	1.00	0.08	2188256	2073102	5.26
50054	1.6	1.7	0.99	0.09	2279388	2133669	6.39
50005_10_50008_90	1.9	1.7	0.98	0.14	2410422	1927536	20.03
50022_10_50016_90	1.9	1.8	0.98	0.13	2465573	2059733	16.46
50027_10_50053_90	1.9	1.7	0.98	0.13	2332200	1943566	16.66
50054_10_50027_90	1.9	1.7	0.98	0.14	2408983	1967530	18.33
50005_20_50008_80	2.2	2.0	0.92	0.18	2781638	1798329	35.35
50022_20_50016_80	2.4	2.1	0.92	0.18	2902550	1863518	35.80
50027_20_50053_80	2.6	2.0	0.92	0.18	2846513	1750277	38.51
50054_20_50027_80	2.4	2.1	0.92	0.18	2894513	1814605	37.31
50005_30_50008_70	3.0	2.6	0.87	0.20	3235816	1613056	50.15
50022_30_50016_70	3.2	2.8	0.88	0.20	3322283	1676965	49.52
50027_30_50053_70	3.0	2.7	0.88	0.20	3254114	1608676	50.56
50054_30_50027_70	3.1	2.8	0.87	0.20	3383218	1618203	52.17
50005_50_50008_50	3.9	2.7	0.82	0.21	3712528	1435999	61.32
50022_50_50016_50	4.1	2.8	0.81	0.21	3798084	1447188	61.90
50027_50_50053_50	3.5	2.6	0.81	0.21	3709487	1425530	61.57
50054_50_50027_50	3.7	2.8	0.82	0.21	3789406	1442491	61.93

дован показатель Call Rate с минимальным пороговым значением 0.95. Call Rate 10%-ных смесей был близок к отметке 0.98, а образцов с большей долей примеси еще ниже. Таким образом, нижнее пороговое значение Call Rate 0.99 позволит выявить образцы с 10% примесей и более.

Результаты работы программы CrossCheck на образцах WGS

При сравнении программой CrossCheck VCF-файла самого с собой или с VCF-файлом, полученным для того же донора биоматериала при повторном секвенировании, значение LOD принимает значения около 5200 (Приложение, табл. 5). При сравнении разных файлов между собой среднее значение LOD составляет –5000. Для WGS данных проекта “1000 геномов” значения LOD несколько различались: так, при сравнении VCF-файла самого с собой LOD составлял в среднем 7594 ± 73, а при сравнении VCF-файлов, полученных для разных образцов, значения LOD имели среднее –12198 ± 248 (данные проекта “1000 геномов”, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/

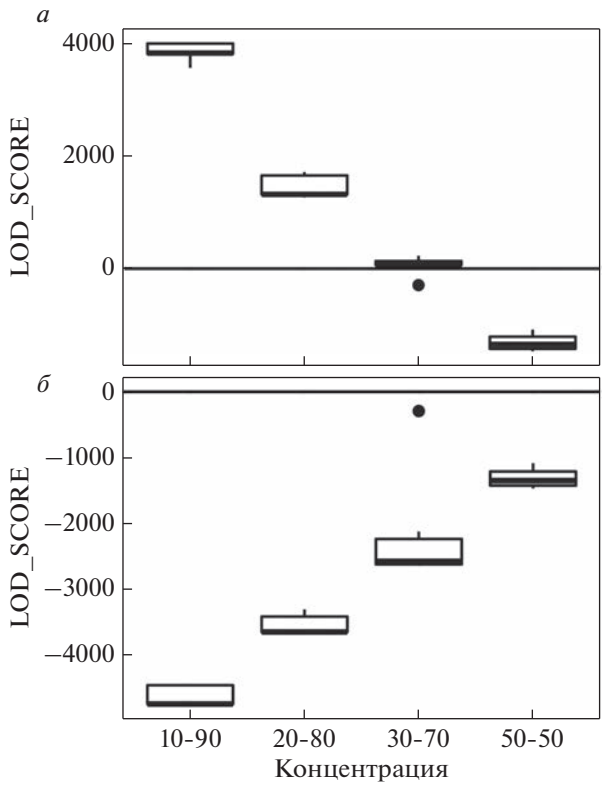

Рис. 1. Значение LOD Score в зависимости от концентрации смеси (процента контаминации) для результатов WGS, полученное при сравнении пар образцов, где один – "чистый" образец, второй – это смесь, где "чистый" образец выступает в роли мажорного (*а*) и минорного (*б*) компонентов в разном процентном соотношении.

data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/). То есть значения LOD имеют узкий диапазон, что позволяет акцентировать внимание на парах образцов с выбивающимся LOD. Значения LOD, полученные для данных проекта "1000 геномов", больше LOD, полученных на наших данных, за счет того, что в файле от проекта "1000 геномов" представлены SNPs, прошедшие фильтрацию, и набор SNPs у всех образцов получился одинаковым. В то время как в наших данных представлены все SNPs, обнаруженные в каждом конкретном образце.

Кроме того, мы дополнительно проверили, что CrossCheck не дает ложноположительных результатов, например при сравнении родственников первого колена LOD находится в диапазоне от −3800 до −1700.

В случае со смесями значение LOD становится ниже. У пар, состоящих из чистого образца и смеси того же образца и 10% контаминирующего агента, значение LOD снижается до 3500; если контаминирующего агента в смеси 20%, то LOD снижается до 1200–1700 (рис. 1,*а*). Для 30%-ных смесей значения LOD попадают в диапазон от −300 до 200, т.е. такие пары могут быть определены и как разные (LOD < −5), и как дубли (если LOD > 5), и как "сомнительные" (−5 < LOD < 5). Пары чистого образца и 50%-ных смесей однозначно помечаются как разные. Пары, состоящие из чистого образца и смеси, где "чистый" образец является минорным компонентом, имеют LOD ниже −1000, т.е. помечаются как разные образцы (рис. 1,*б*).

Кроме того, в ходе выполнения исследования мы попарно сравнили смеси одних и тех же образцов с разным процентным соотношением компонентов (рис. 2,*а*). Чем больше схож процентный состав двух смесей, тем выше LOD. При сравнении файла самого с собой вне зависимости от процента примеси LOD получается около 5000 или выше, т.е. такой же, как у чистых образцов. Чуть меньшие значения (4500) принимает LOD при сравнении 30%-ных и 50%-ных смесей между собой. Во всех остальных случаях LOD ниже 4000. Таким образом, заниженное значение LOD может быть дополнительным признаком контаминации. Смеси с содержанием контаминирующего агента 30% и более при сравнении с образцом, соответствующим мажорному компоненту, могут не быть определены как поступившие от одного донора биоматериала (LOD близок к 0, а при 50% – ниже 0).

При сравнении смесей, состоящих из разных образцов, LOD возрастает вместе с долей контаминации (рис. 2,*б*). Однако LOD не превышает 0, т.е. даже 50%-ные смеси будут помечаться как разные. Выбросы выше 0 (рис. 2,*б*) обусловлены тем, что один и тот же образец использовался как основной компонент в одной смеси и как минорный компонент в другой смеси. В этом случае LOD не превышает 1000 и остается достаточно низким по сравнению со значением 5000, характерным для чистых образцов. Выбросы, расположенные ниже 0, обусловлены тем, что среди доноров образцов есть родственники первого колена.

Результаты работы программы CrossCheck на образцах CoreExome

Технология определения однонуклеотидных вариантов Illumina microarray BeadArray technology CoreExome позволяет провести генотипирование по 567 218 точкам, а не по всему геному, как в случае с WGS, причем только 2998 из них входят в референсный файл CrossCheck. Это приводит к тому, что LOD при сравнении файлов, полученных с помощью чипов CE, будет ниже, чем в случае WGS.

При сравнении файла самого с собой или с дублем среднее значение LOD составляет 1000, а для пары разных файлов −700 (Приложение,

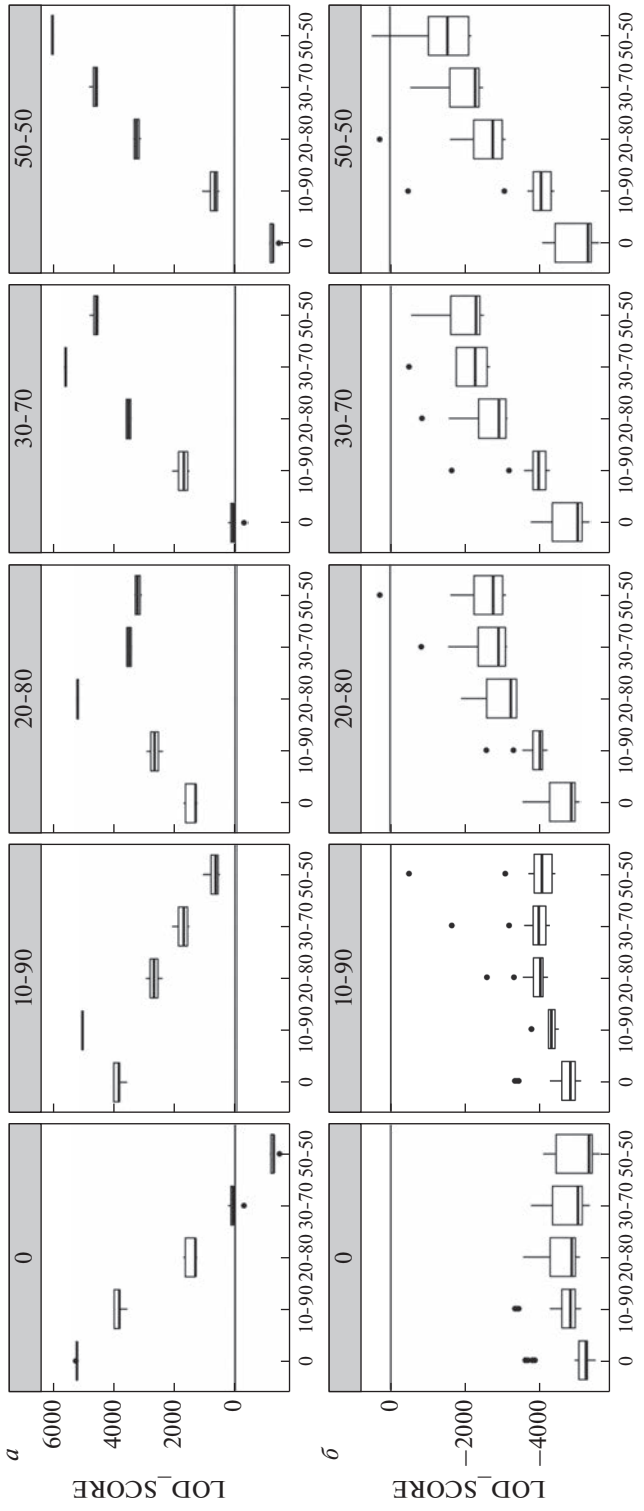

Рис. 2. Значение LOD при сравнении WGS смесей одних и тех же образцов между собой (*а*) и образцов из разных “партий” (*б*). По оси абсцисс отмечена концентрация (процент контаминации) одного из образцов из пары, в боксах сверху – процент контаминации второго образца из пары.

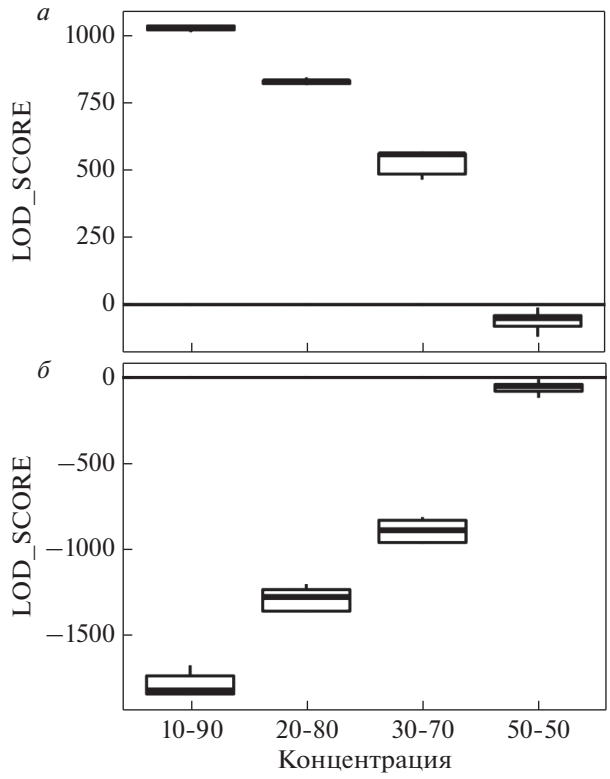

Рис. 3. Значение LOD Score в зависимости от концентрации смеси (процента контаминации) для результатов CE, полученное при сравнении пар образцов, где один – “чистый” образец, второй – это смесь, где “чистый” образец выступает в роли мажорного (*а*) и минорного (*б*) компонентов в разном процентном соотношении.

табл. 6). Также мы установили характерные значения LOD для данных проекта “1000 геномов”. Из WGS-данных проекта “1000 геномов” отфильтровали только SNPs, входящие в CE, и сравнили полученные файлы с помощью программы CrossCheck. При сравнении файла самого с собой LOD в среднем находится на отметке 1092 ± 16, а при сравнении VCF-файлов, полученных для разных образцов, LOD равен −1980 ± 56.

Для пары образцов, генотипированных с помощью CE, LOD составляет около 1000 в том случае, если доля примеси не превышает 10%. Во всех остальных случаях LOD будет ниже 900 (рис. 3,*а*; 4,*а*). Это позволяет распознать образцы, которые контаминированы или по какой-либо другой причине имеют низкое качество. Пары, включающие образец и смесь с этим же образцом в качестве минорного компонента, имеют LOD ниже 0 (рис. 3,*б*). Для пары смесей разных образцов LOD возрастает вместе с долей контаминации (рис. 4,*б*). Однако LOD не превышает 0, т.е. даже сильно контаминированные образцы будут помечаться как разные.

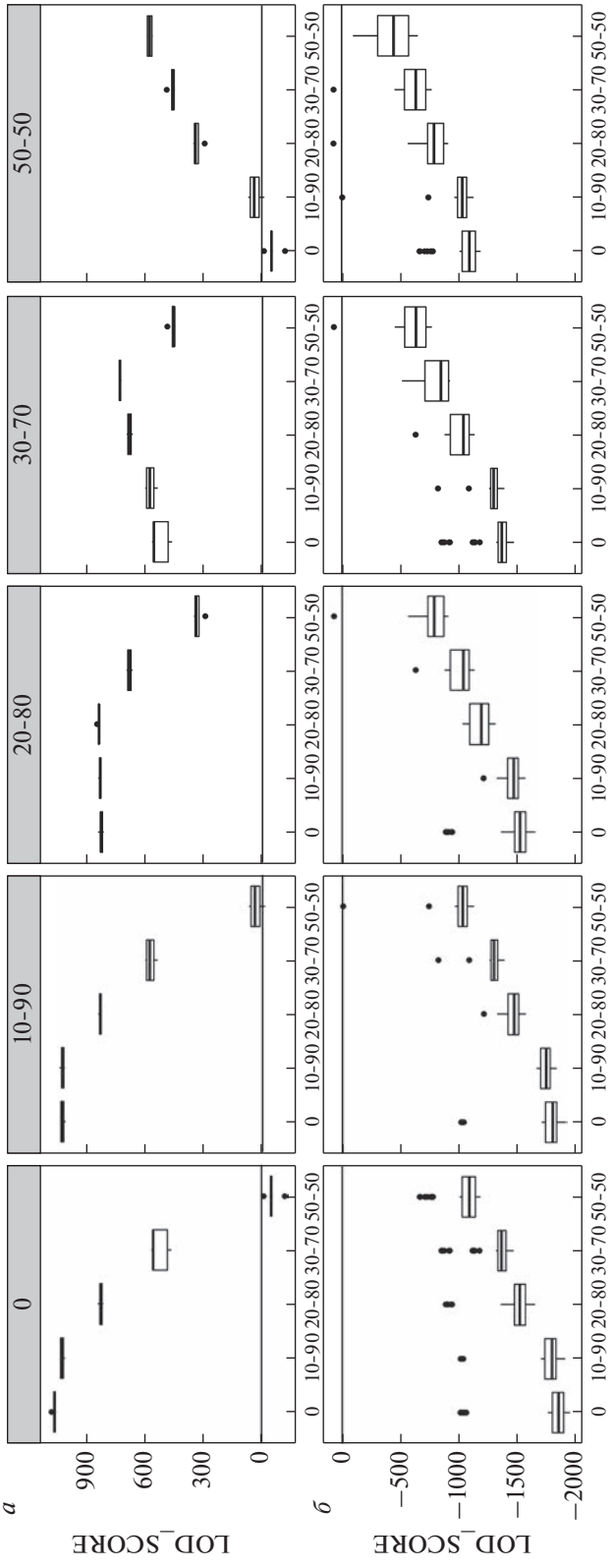

Рис. 4. Значение LOD при сравнении CE смесей одних и тех же образцов между собой (*а*) и образцов из разных “партий” (*б*). По оси абсцисс отмечена концентрация (процент контаминации) одного из образцов из пары, в боксах сверху – процент контаминации второго образца из пары.

Работа CrossCheck с парой образцов CoreExome–WGS

Если в рамках эксперимента один и тот же образец был генотипирован с помощью CE и WGS, то необходимо убедиться, что не возникло путаницы и действительно результаты CE и WGS принадлежат одному и тому же образцу. Для этих целей также можно воспользоваться программой CrossCheck. Сравнение файлов CE и WGS происходит по 2998 точкам. Значения LOD для идентичных образцов составляет примерно 650, для пары разных образцов примерно −1500 (Приложение, табл. 7).

В отличие от сравнения результатов, полученных с помощью одной и той же технологии, в данном случае имеет значение какой из образцов (CE или WGS) контаминирован. Если контаминирован образец WGS, то обнаружить соответствующий ему образец на CE можно только при контаминации не выше 10% (рис. 5,*а*, *б*). Если же контаминирован образец CE, то обнаружить соответствующий ему WGS можно даже при контаминации в 30% (рис. 5,*в*, *г*). А при контаминации CE образца на 50% LOD будет принимать значения близкие к 0, т.е. возможны ситуации определения пары образцов как разных (LOD < −5), как дублей (если LOD > 5) или как “сомнительных” (−5 < LOD < 5).

В случае контаминации обеих проб образца (WGS до 10% и CE до 20%) LOD будет держаться на уровне 600 (рис. 6,*а*). При более высоком уровне контаминации LOD для пары, где хоть одна из проб контаминирована, будет принимать значения значительно ниже 600, т.е. по значению LOD можно предположить наличие контаминации, как минимум, в одном из образцов.

Смеси, состоящие из пар разных образцов, при сравнении друг с другом имеют LOD ниже 0, т.е. всегда однозначно помечаются как разные (рис. 6,*б*).

ОБСУЖДЕНИЕ

В рамках настоящего исследования были проанализированы метрики качества “чистых” и контаминированных образцов ДНК человека. Впервые показано, что метрики качества, рассчитанные по файлам с прочтениями (FASTQ) и выравниванием (BAM), между “чистыми” и контаминированными образцами не различались. Однако метрики, рассчитанные на базе файлов, содержащих информацию о SNPs (VCF-файлы), позволяют отличать контаминированные и “чистые” образцы друг от друга.

У всех способов контроля качества, использованных в данной работе, есть свой диапазон допустимых значений, именно поэтому целесообразно использовать несколько методов одновременно. Для WGS набор методов может включать в себя

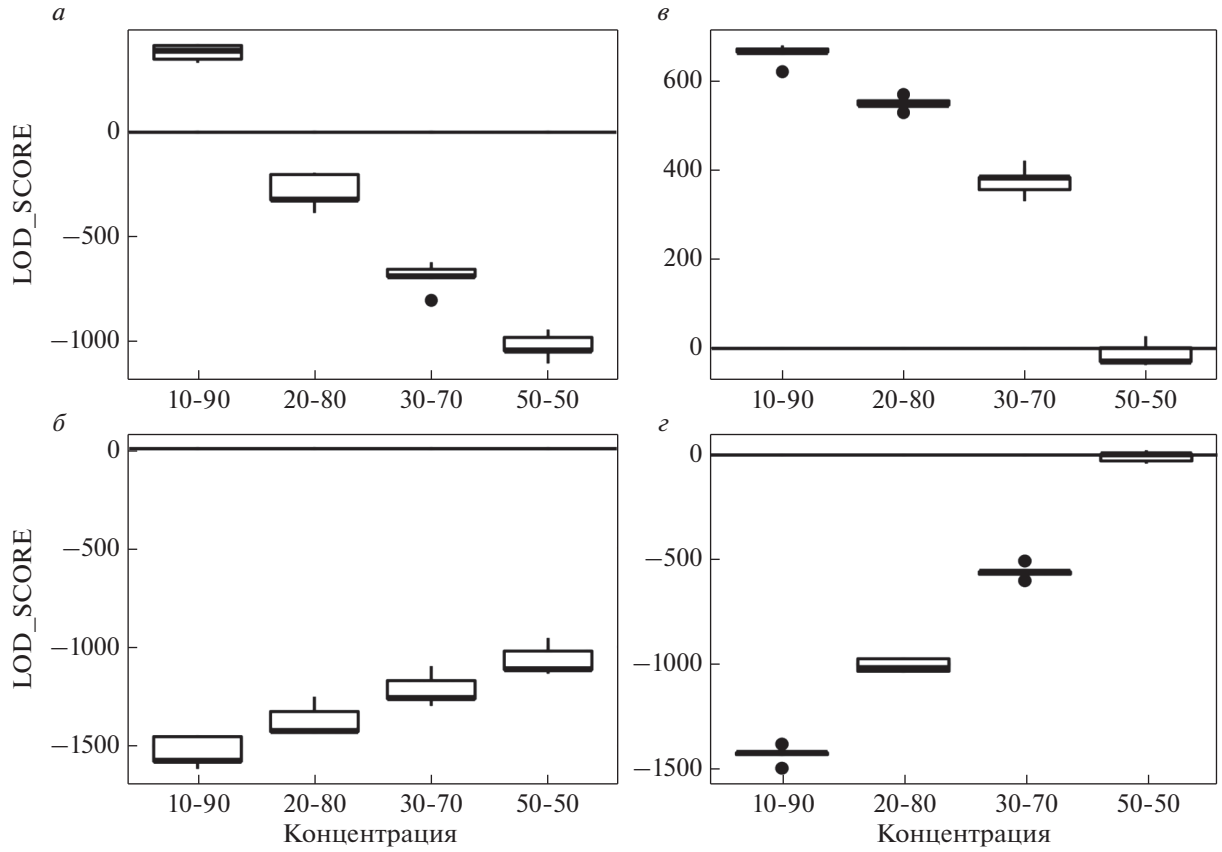

Рис. 5. Значение LOD Score, полученное при сравнении: *а* – “чистого” CE и контаминированного WGS для пары идентичных образцов; *б* – “чистого” CE и смеси, где “чистый” образец выступает в роли минорного компонента в разном процентном соотношении; *в* – “чистого” WGS и контаминированного CE для пары идентичных образцов; *г* – “чистого” WGS и смеси, где “чистый” образец выступает в роли минорного компонента в разном процентном соотношении.

определение процента гетерозиготных вариантов с AR вне диапазона $0.5 \pm 1.96 \times 0.09$ и стандартного отклонения AR, расчет Het/Hom и использование программы CrossCheck. Результаты, полученные в данной работе, подтверждают предыдущие результаты, опубликованные Dallavilla и соавт.: расчет AR позволяет выявлять контаминацию 10% и более [1]. Впервые рассчитана метрика Het/Hom для образцов с разной степенью контаминации. Повышенное значение Het/Hom может трактоваться только как косвенное свидетельство контаминации и только совместно с информацией об этнической принадлежности донора биоматериала. В настоящей работе проверили корректность работы программы CrossCheck с контаминированными образцами. Программа CrossCheck позволяет выявлять пары образцов биоматериала, поступивших от одного и того же донора и генотипированных с помощью CE и WGS. Впервые показано, что метрика схожести файлов, рассчитанная программой CrossCheck, зависит от степени контаминации образца. При этом разные контаминированные образцы при сравнении друг с другом имеют LOD ниже 0, т.е. всегда однозначно помечаются как разные. Для того чтобы иметь возможность использовать LOD как дополнительную метрику качества образца, стоит оценить характерные для конкретного набора данных значения LOD для пары проб, поступивших от одного донора биоматериала, и для пары разных образцов.

Поскольку принципы генотипирования и секвенирования любых других диплоидных организмов не отличаются от подходов, используемых для расшифровки нуклеотидной последовательности ДНК человека, описанные в данной работе методы можно применять и для их контроля качества.

Исследование не имело спонсорской поддержки.

Все процедуры, выполненные в исследовании с участием людей, соответствуют этическим стандартам институционального и/или национального комитета по исследовательской этике и Хельсинкской декларации 1964 г. и ее последующим изменениям или сопоставимым нормам этики.

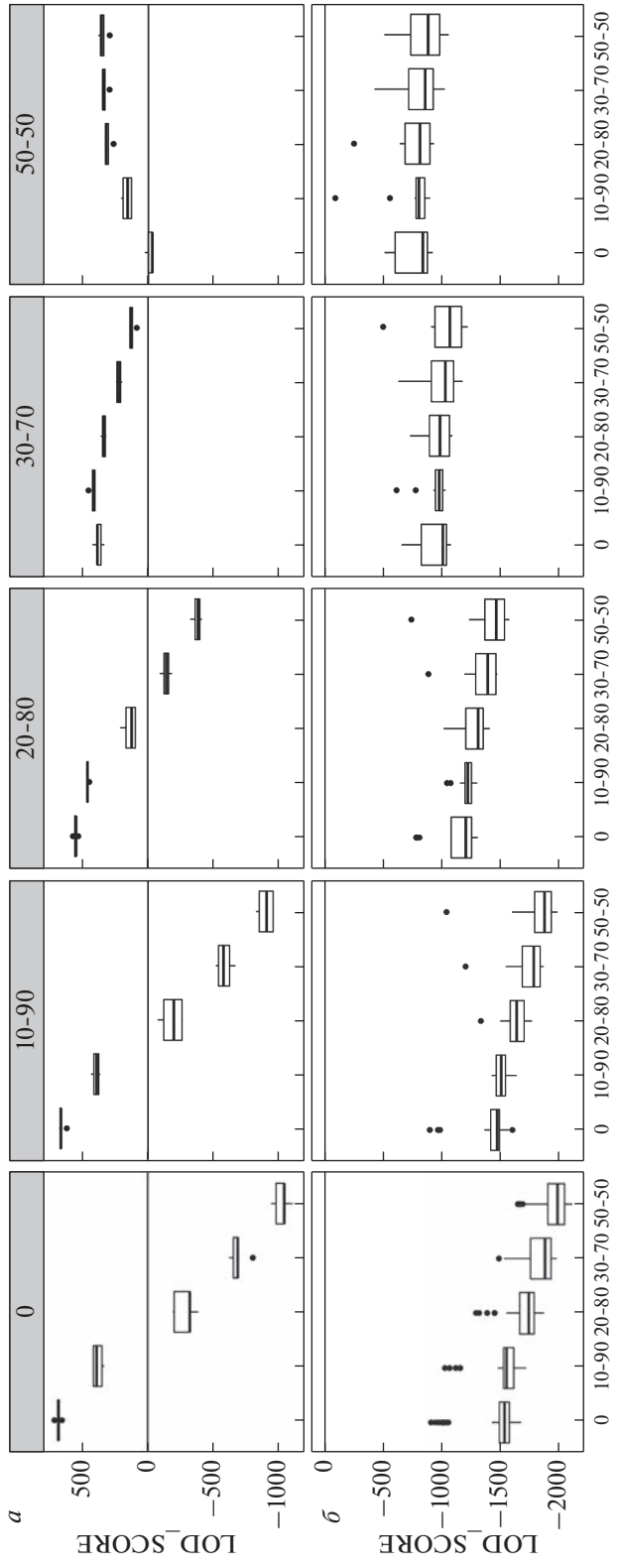

Рис. 6. Значение LOD при сравнении CE и WGS смесей одних и тех же образцов между собой (*а*) и контаминированных образцов из разных “партий” (*б*). По оси абсцисс отмечена концентрация (процент контаминации) WGS, в боксах сверху – процент контаминации CE.

От каждого из включенных в исследование участников было получено информированное добровольное согласие.

Авторы заявляют, что у них нет конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. *Dallavilla T., Marceddu G., Casadei A. et al.* A fast, reliable and easy method to detect within-species DNA contamination // Acta Bio-Medica Atenei Parm. 2020. V. 91. № 13-S. https://doi.org/10.23750/abm.v91i13-S.10531
2. *Wang J., Raskin L., Samuels D.C. et al.* Genome measures used for quality control are dependent on gene function and ancestry // Bioinformatics. 2015. V. 31. № 3. P. 318–323. https://doi.org/10.1093/bioinformatics/btu668
3. *Javed N., Farjoun Y., Fennell T.J. et al.* Detecting sample swaps in diverse NGS data types using linkage disequilibrium // Nat. Commun.2020. V. 11. № 1. P. 3697. https://doi.org/10.1038/s41467-020-17453-5
4. *Miller N.A., Farrow E.G., Gibson M. et al.* A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases // Genome Med. 2015. V. 7. № 1. P. 100. https://doi.org/10.1186/s13073-015-0221-8
5. *Kim S., Scheffler K., Halpern A.L. et al.* Strelka2: Fast and accurate calling of germline and somatic variants // Nat. Methods. 2018. V. 15. № 8. P. 591–594. https://doi.org/10.1038/s41592-018-0051-x
6. *Danecek P., Bonfield J.K., Liddle J. et al.* Twelve years of SAMtools and BCFtools // GigaScience. 2021. V. 10. № 2. https://doi.org/10.1093/gigascience/giab008
7. *Zhao H., Sun Z., Wang J. et al.* CrossMap: A versatile tool for coordinate conversion between genome assemblies // Bioinforma. Oxf. Engl. 2014. V. 30. № 7. P. 1006–1007. https://doi.org/10.1093/bioinformatics/btt730

Experimental Evaluation of the Possibility to Detect Cross-Contaminated DNA Samples Based on Genetic Data

N. V. Feliz[a], *, K. S. Grammatikati[a], S. I. Mitrofanov[a], P. A. Grebnev[a], K. D. Konureeva[a], E. D. Maralova[a], M. V. Erokhina[a], T. A. Shpakova[a], P. G. Kazakova[a], Yu. N. Akhmerova[a], A. A. Mkrtchian[a], E. A. Snigir[a], V. S. Yudin[a], A. A. Keskinov[a], S. M. Yudin[a], and V. I. Skvortsova[b]

[a] Federal State Budgetary Insntitution "Centre for Strategic Planning and Management of Biomedical Health Risks" of the Federal Medical Biological Agency, Moscow, 119121 Russia

[b] The Federal Medical Biological Agency, Moscow, 123182 Russia

**e-mail: feliz08nv@gmail.com*

The problems of cross-contamination and swap samples are extremely relevant during large-scale genetic studies. In this study several approaches of detecting cross-contaminated DNA samples were checked: the ratio of reads per reference and alternative allele (allele ratio, AR), the amount of heterozygos to homozygous variants ratio, the CallRate value for the DNA microarrays data, the Picard CrosscheckFingerprints (CrossCheck) program. Contaminated samples (mixtures) were created by mixing ordinary "pure" DNA samples in different ratios. Samples' quality parameters were analyzed after whole genome sequencing and genotyping with the Illumina microarray BeadArray technology CoreExome (CE) DNA microarray. It has been experimentally established that all of these approaches can be used to detect genotyping errors associated with sample contamination.

Keywords: whole genome sequencing, contamination, microarray BeadArray technology, quality control.