БИОИНФОРМАТИКА

BIOINFORMATICS

УДК 519.23
DOI: 10.37661/1816-0301-2025-22-3-45-58

Оригинальная статья
Original Article

Алгоритм выбора референсных микроРНК при классификации биологических процессов

О. В. Красько[1]✉, С. В. Якубовский[2], В. Н. Кипень[3]

[1]*Объединенный институт проблем информатики Национальной академии наук Беларуси, ул. Сурганова, 6, Минск, 220012, Беларусь*
✉*E-mail: krasko@newman.bas-net.by*

[2]*Белорусский государственный медицинский университет, пр. Дзержинского, 83, Минск, 220083, Беларусь*
E-mail: yakub-2003@yandex.by

[3]*Институт генетики и цитологии Национальной академии наук Беларуси, ул. Академическая, 27, Минск, 220072, Беларусь*
E-mail: v.kipen@igc.by

Аннотация

Цели. Целью исследования является разработка алгоритма выбора референсных микроРНК с учетом их взаимосвязи с тем, чтобы классифицировать группы образцов при изучении различных биологических процессов.

Методы. Использовались методы линейной алгебры, анализа главных компонент, статистических моделей бинарной регрессии, оценки производительности моделей.

Результаты. Разработан новый алгоритм MDSeek, который предлагает выбор референсных микроРНК для нормализации данных количественной полимеразной цепной реакции с целью последующего использования нормализованных данных для задач классификации. Оценка результатов работы алгоритма для задачи классификации свидетельствует о его более высокой эффективности по сравнению с известными подходами к нормализации результатов полимеразной цепной реакции.

Заключение. В настоящей работе предложен оригинальный алгоритм MDSeek, предназначенный для выбора референсных микроРНК с целью нормализации результатов полимеразной цепной реакции и позволяющий изучать изменения экспрессии микроРНК при сравнении различных биологических процессов. После применения MDSeek на опытном наборе образцов нормализованные данные использовались для задач классификации, метрики производительности были лучше по сравнению с другими алгоритмами.

Ключевые слова: микроРНК, нормализация, полимеразная цепная реакция, классификация, расстояние Махаланобиса, производительность моделей

Для цитирования. Красько, О. В. Алгоритм выбора референсных микроРНК при классификации биологических процессов / О. В. Красько, С. В. Якубовский, В. Н. Кипень // Информатика. – 2025. – Т. 22, № 3. – С. 45–58. – DOI: 10.37661/1816-0301-2025-22-3-45-58.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 17.07.2025
Подписана в печать | Accepted 31.07.2025
Опубликована | Published 30.09.2025

Algorithm for selecting reference microRNAs in biological processes classification

Olga V. Krasko[1]✉, Siarhei U. Yakubouski[2], Viachaslau N. Kipen[3]

[1]The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, st. Surganova, 6, Minsk, 220012, Belarus
✉E-mail: krasko@newman.bas-net.by

[2]Belarusian State Medical University, av. Dzerzhinski, 83, Minsk, 220116, Belarus
E-mail: yakub-2003@yandex.by

[3]The Institute of Genetics and Cytology of the National Academy of Sciences of Belarus, st. Akademicheskaya, 27, Minsk, 220072, Belarus
E-mail: v.kipen@igc.by

Abstract

Objectives. The algorithm for selection of reference microRNA taking into account their biological features for classification of pathologies.

Development of an algorithm for selecting microRNAs with regard to their interconnection for samples classification in the various biological processes

Methods. Methods of linear algebra, principal component analysis, statistical binary regression models, and model performance metrics were used.

Results. A new algorithm, MDSeek, has been developed that proposes a selection of reference microRNA for the normalization quantitative polymerase chain reaction results taking into account their coexpression. MDSeek demonstrates higher performance metrics compared to known reference gene selection approaches for the subsequent classification tasks.

Conclusion. An original MDSeek algorithm for selecting reference microRNAs for normalization results of polymerase chain reaction is suggested. It takes into account changes in microRNA expression when comparing different biological processes. After applying MDSeek to an experimental set of samples, the normalized data were used for classification tasks, and the performance metrics were better than those of other normalization algorithms.

Keywords: MicroRNA, normalization, polymerase chain reaction, classification, Mahalanobis distance, model performance

For citation. Krasko O. V., Yakubouski S. U., Kipen V. N. *Algorithm for selecting reference microRNAs in biological processes classification*. Informatika [*Informatics*], 2025, vol. 22, no. 3, pp. 45–58 (In Russ.). DOI: 10.37661/1816-0301-2025-22-3-45-58.

Conflict of interest. The authors declare of no conflict of interest.

Введение. Нормализация в количественной полимеразной цепной реакции (quantitative polymerase chain reaction, qPCR) является критически важным этапом предобработки данных для достижения точной и надежной количественной оценки уровней экспрессии, получаемых в результате qPCR в различных биологических образцах.

Нормализация – это процесс корректировки относительных мер экспрессии между образцами для снижения технических отклонений данных qPCR, не имеющих биологических причин (эффективность qPCR, качество образцов и т. д.), что необходимо для получения биологически значимых и воспроизводимых результатов, которые можно использовать для последующего анализа. Поскольку при выполнении qPCR исследуются относительные значения изменения экспрессии – насколько больше или меньше экспрессируются одни микроРНК относительно других, нормализация предусматривает использование референсных показателей (нормализаторов), по отношению которых изучается экспрессия исследуемых микроРНК. Такими норма-

лизаторами служат молекулы микроРНК, выбор которых может базироваться на различных принципах. Необходимость корректной нормализации подчеркивается получением различных, зачастую противоречивых результатов в ряде исследований, посвященных проблеме изучения экспрессии микроРНК. Некорректная нормализация может приводить к неверной оценке или отсутствию выявления имеющихся изменений в профиле экспрессии микроРНК и сопровождается риском ошибочных выводов при сравнении образцов из разных условий или источников, что делает ее фундаментальным аспектом исследований в области молекулярной биологии [1].

Для повышения надежности результатов qPCR были разработаны различные методы нормализации, базирующиеся на пороговом числе циклов PCR[1] (Ct), включая метод ΔCt, квантильную нормализацию и рангово-инвариантную нормализацию набора образцов. Метод ΔCt сравнивает значения порога цикла (Ct) целевых генов со значениями референсных генов. Квантильная и рангово-инвариантная нормализации предлагают более сложные подходы, смягчая технический шум и устраняя зависимость от одного референсного гена соответственно. Вместе с тем в настоящее время не существует единого стандартного подхода к выбору референсной микроРНК, поскольку зависимость от одного показателя может внести значительную ошибку в результаты нормализации [2–4].

Текущие рекомендации предлагают использовать несколько референсных генов, в идеале три или более, чтобы обеспечить надежную нормализацию, которая может учитывать присущую изменчивость различных тканей и экспериментальных условий. Такая практика соответствует рекомендациям MIQE, направленным на стандартизацию методологий qPCR и улучшение воспроизводимости в исследованиях [5].

Применение нормализованных данных qPCR охватывает различные области, включая исследования рака, биологию развития и фармакогеномику, где они облегчают анализ паттернов экспрессии генов, относящихся к болезням и ответам на лечение. Продолжающаяся эволюция стратегий и инструментов нормализации повышает надежность и воспроизводимость результатов qPCR, устраняя сложности, связанные с анализом экспрессии генов [6].

Таким образом, нормализация как процесс предобработки пытается решить проблемы, связанные со следующими факторами:

1) контролем за вариациями, поскольку в экспериментах с использованием qPCR существует множество факторов, которые могут вызвать вариации в результатах, такие как количество исходной РНК, эффективность обратной транскрипции и qPCR, качество образцов и т. д. Нормализация помогает контролировать эти вариации, что делает данные более сопоставимыми;

2) снижением экспериментальных ошибок, поскольку нормализация позволяет учитывать технические вариации между образцами, что повышает точность данных;

3) калибровкой данных, так как нормализация с использованием референсных генов помогает откалибровать результаты экспрессии целевых генов, чтобы снизить технические и биологические вариации;

4) унификацией при сравнительных исследованиях, когда необходимо выявить дифференциальную экспрессию генов.

Многие методы нормализации основаны на предположении, что один или несколько генов конститутивно экспрессируются на почти постоянных уровнях при всех экспериментальных условиях и уровни экспрессии всех генов в образце нормализуются с учетом этого предположения.

1. Основные стратегии выбора референсных генов, основанные на ΔCt. Предполагается, что имеется матрица количественных данных размером $N \times K$, где N – число рассматриваемых генов, K – число образцов исследуемого материала. Каждая ячейка матрицы содержит значение Ct – число циклов PCR, при котором флуоресценция превышает пороговое значение для *n*-го гена в *k*-м образце.

[1]Пороговое число циклов (ΔCt) – это число циклов PCR, при котором флуоресценция превышает пороговое значение.

Рассмотрим следующие алгоритмы:

1. Алгоритм BestKeeper [7] предназначен для выбора стабильных референсных генов для использования в исследованиях экспрессии генов с помощью qPCR. Основные шаги алгоритма:

1. Расчет стандартного отклонения: для каждого кандидата в референсные гены вычисляется стандартное отклонение (SD) и среднее значение $(Mean)$ Ct значений во всех образцах.

2. Оценка стабильности: стабильность каждого кандидата в референсные гены оценивается по соотношению стандартного отклонения к среднему значению $(SD/Mean)$. Наименьшие значения этого соотношения указывают на наибольшую стабильность.

3. Выбор референсных генов: из всех кандидатов выбираются те, которые имеют наименьшие значения соотношения $(SD/Mean)$, т. е. наиболее стабильные.

2. Алгоритм geNorm [1] также предназначен для выбора наиболее стабильных референсных генов для нормализации данных qPCR. Он основан на оценке стабильности генов путем анализа их выраженности в различных образцах. Основные шаги алгоритма:

1. Расчет коэффициента стабильности M для каждого гена, основанного на среднегеометрической парной вариации между генами. Гены с наименьшим значением M считаются наиболее стабильными.

2. Постепенное исключение менее стабильных генов: на каждом шаге исключается ген с наибольшим значением M, и процесс повторяется до тех пор, пока не останутся два гена. Таким образом, в каждом цикле исключается наименее стабильный ген.

3. Определение оптимального числа референсных генов путем расчета коэффициента вариации (V) для последовательных пар генов. Если коэффициент вариации V между генами n и $n+1$ ниже заданного порогового значения (например, 0,1), считается, что добавление дополнительного гена не улучшит точность нормализации.

4. Расчет нормализационного фактора (NF) для каждого образца как среднегеометрического значения экспрессии выбранных референсных генов в данном образце на основе выбранных стабильных генов.

3. Алгоритм NormFinder (два варианта: с группами и без групп) [8] – это статистический метод для выбора наиболее стабильных референсных генов с целью нормализации данных qPCR. Основные шаги алгоритма:

1. Расчет внутригрупповой и межгрупповой вариации. Внутригрупповая вариация – это вариация Ct значений одного гена внутри одной группы образцов, межгрупповая вариация – вариация средних Ct значений одного гена между разными группами образцов.

2. Расчет комбинированной стабильности для каждого гена. Алгоритм вычисляет комбинированную меру стабильности для каждого гена, объединяя внутригрупповую и межгрупповую вариации. Гены с наименьшей комбинированной мерой стабильности считаются наиболее стабильными и подходят для использования в качестве референсных.

3. Ранжирование генов: все кандидаты ранжируются по их комбинированной стабильности. Те, у кого эта мера наименьшая, считаются наиболее стабильными и рекомендованы для использования в качестве референсных.

Преимущества алгоритма:

– учет вариаций. Алгоритм NormFinder учитывает как внутригрупповую, так и межгрупповую вариации, что делает его особенно полезным при анализе сложных данных;

– различное число групп. Алгоритм NormFinder может быть применен как к набору образцов с одним классом, так и к данным с несколькими классами.

4. Алгоритм NormiRAZOR [9] является математической комбинацией всех перечисленных выше алгоритмов. Он осуществляет полный перебор сочетаний $C_n^m = \frac{N!}{(N-M)!\times M!}$, где N – число рассматриваемых генов, M – число генов, которые выбраны как референсные. Для каждого сочетания рассчитываются меры стабильности для четырех вариантов: BestKeeper,

geNorm, NormFinder (два варианта: с классами и без них). Далее проводится ранжирование по всем сочетаниям по каждому алгоритму отдельно и рассчитываются ранги всех сочетаний по каждому алгоритму отдельно. Ранги приводятся к единичному диапазону. Для каждого сочетания определяется комбинированный ранг как среднее четырех рангов (единичного интервала). Окончательное ранжирование сочетаний проводится по усредненному рангу. Выбор сочетания, которое будет набором референсных генов, остается за исследователем.

Разные стратегии могут значительно исказить способность нахождения различий между группами образцов, связанных с рядом заболеваний (формами заболевания), и, безусловно, являются существенным недостатком при сравнении уровней экспрессии между различными исследованиями. В дополнение к стандартизации сбора и обработки образцов использование конкретной процедуры нормализации обеспечит лучшую воспроизводимость, а определение эндогенных эталонных генов расширит возможности проведения крупномасштабных анализов на выбранных маркерах [10].

2. Особенности анализа данных экспрессии микроРНК. МикроРНК – это короткие (18–24 нуклеотида) некодирующие молекулы РНК, играющие важную роль в регуляции экспрессии генов. Показано, что мишенями микроРНК являются от 30 до 60 % генов человека, кодирующих разнообразные белки [11]. Было установлено, что молекулы данного типа экспрессируются в опухолях различного генеза, их аберрантная экспрессия играет важную роль в пролиферации, дифференцировке, инвазии, миграции и апоптозе опухолевых клеток [12]. Усиленная экспрессия микроРНК может быть связана с возникновением злокачественной опухоли, и такие микроРНК являются онкогенными. Подавление экспрессии микроРНК зачастую подавляет и развитие опухоли. Описаны также микроРНК с опухоль-супрессорной активностью. Показано, что функция каждой микроРНК может быть ткане- и контекст-специфичной [13].

МикроРНК являются ключевыми игроками в сложных биологических процессах, нарушения их экспрессии связаны, помимо опухолей, с развитием многих других заболеваний, включая сердечно-сосудистые, нейродегенеративные и инфекционные. Это обусловливает возможность использования микроРНК в качестве диагностических и прогностических биомаркеров широкого спектра заболеваний. Изучение микроРНК помогает понять механизмы развития различных заболеваний и разработать новые методы их диагностики и лечения.

Особенностью экспрессии микроРНК являются ее высокая вариабельность и корреляция между изучаемыми микроРНК.

Биологическая вариабельность обусловлена гетерогенностью клеток. В различных клеточных типах и тканях уровни экспрессии микроРНК могут существенно различаться. Это усложняет интерпретацию результатов, поскольку они могут отражать не только интересующее исследователей состояние, но и общую клеточную гетерогенность. Также уровень микроРНК может изменяться в зависимости от физиологических состояний организма, таких как стресс, воспаление, диета, возраст и т. д.

Корреляция определена коэкспрессией микроРНК, в некоторых случаях несколько микроРНК могут коэкспрессироваться и оказывать схожие эффекты на экспрессию генов. Это может затруднять выделение вкладов отдельных микроРНК в общий результат [14].

В отличие от генов, которые могут быть конститутивно обусловлены, микроРНК являются гибким биологическим инструментом в регуляции экспрессии самих генов. Таким образом, изучение микроРНК является контекстуальным и в первую очередь зависит от конкретной решаемой задачи. Одной из таких задач является задача классификации, и в настоящей работе предлагается включить в процесс выбора референсных микроРНК метрики, связанные с различием классов (изучаемых вариантов опухолевого процесса) изначально, для того чтобы комплексно решать задачу нормализации совместно с задачей классификации. Такие алгоритмы, как BestKeeper и geNorm, изначально не включают классы. Алгоритм NormFinder включает как внутригрупповую, так и межгрупповую вариацию, но не учитывает корреляцию между *Ct* различных микроРНК, что характерно и важно для биологических процессов.

3. Предлагаемый алгоритм (MDSeek). При выборе референсных микроРНК предлагается рассмотреть такую метрику, как расстояние Махаланобиса [15, 16], основной смысл использования которой – учет дисперсий и ковариаций пространства признаков. Также особенностью

предлагаемого метода выбора является возможность анализа пространства признаков (исследуемых микроРНК), который превышает размер образцов определенного класса.

Пусть имеется N рассматриваемых микроРНК (признаков), которые получены для K образцов тканей. Образцы могут принадлежать G различным классам (изучаемые патологии). Составляются интересующие сочетания или полный перебор сочетаний, общее число $V = C_n^m = \frac{N!}{(N-M)! \times M!}$, где N – число рассматриваемых признаков, M – число признаков, которые выбраны как референсные.

При рассмотрении конкретного сочетания референсных микроРНК $v \in V$ выполняются алгоритмы, состоящие из следующих шагов:

1. Для получения матрицы нормализованных значений $\|\Delta Ct\|_v$ для каждого образца рассчитывается вектор ΔCt длиной $(N-M)$ как разность между значением каждого признака и сочетанием референсных микроРНК. Комбинацией значений референсных микроРНК является их среднее (арифметическое или геометрическое [7]).

2. Выделяются образцы только одного из классов из $\|\Delta Ct\|_v$ и рассчитывается матрица ковариаций признаков $\mathrm{cov}(g,v)$ каждого класса, $g = 1,\ldots,G$. Если хотя бы одна из матриц необратима, то используем метод усадки (shrinkage) ковариационной матрицы [17] для всех классов, который применим в задачах с небольшим количеством образцов и большим количеством признаков (микроРНК), и получаем $\mathrm{cov}_*^{-1}(g,v)$. Если все матрицы обратимы, то далее используем $\mathrm{cov}_*^{-1}(g,v) = \mathrm{cov}^{-1}(g,v)$.

3. Рассчитывается центроид $centroid(g,v)$ признаков каждого класса $g = 1,\ldots,G$ как среднее каждого признака в классе.

4. Число пар сравниваемых классов определяем как число сочетаний $C_G^2 = \frac{G!}{(G-2)! \times 2!}$.

5. Рассчитывается общая ковариационная матрица двух классов как

$$\mathrm{cov}(p,q,v) = \left(\left(n_p - 1\right) \times \mathrm{cov}(p,v) + \left(n_q - 1\right) \times \mathrm{cov}(q,v)\right) \Big/ \left(n_p + n_q - 2\right),$$

где n_p, n_q – число образцов в соответствующих классах $p,q \in G$.

6. Расстояние Махаланобиса определяется для центроидов пары классов

$$MD_{pq}^v = MD\left(centroid(p,v), centroid(q,v), \mathrm{cov}^{-1}(p,q,v)\right) =$$
$$= \sqrt{\left(centroid(p,v) - centroid(q,v)\right) \times \mathrm{cov}^{-1}(p,q,v) \times \left(centroid(p,v) - centroid(q,v)\right)^T}.$$

7. Для выбранного сочетания v референсных микроРНК определяется сумма расстояний Махаланобиса по всем парам класса, повторяя пп. 5, 6 для каждой пары классов.

8. Пп. 1–7 повторяются для каждого сочетания $v \in V$ референсных микроРНК.

9. Находится сочетание микроРНК (признаков), для которых сумма расстояний Махаланобиса по всем парам сравниваемых классов максимальна. Данное сочетание будет использовано как референсное.

4. Сравнительные результаты решения классификационной задачи при различных алгоритмах нормализации. Для проверки эффективности классификации предложенного алгоритма был использован набор образцов, содержащий данные о профиле экспрессии микроРНК и полученный на материале 299 гистологических образцов ткани щитовидной железы

(169 образцов злокачественных опухолей и 130 образцов доброкачественных опухолей). Все образцы представляли собой операционный материал пациентов, проходивших лечение в учреждении здравоохранения «Минский городской клинический онкологический центр» в период 2021–2023 гг. Молекулярно-генетические исследования выполнены на базе Института генетики и цитологии Национальной академии наук Беларуси. Исследование одобрено Комитетом по биомедицинской этике учреждения образования «Белорусский государственный медицинский университет» Министерства здравоохранения Республики Беларусь (протокол № 9 от 23.03.2022). Все данные, полученные из медицинских записей, были анонимизированы.

Методом qPCR были получены пороговые значения Ct 18 микроРНК для каждого образца. На основании проведенного анализа научных публикаций об информативности микроРНК в диагностике злокачественных опухолей щитовидной железы были отобраны следующие микроРНК (табл. 1).

Таблица 1
Перечень исследованных микроРНК

Table 1
List of microRNA examined

МикроРНК *MicroRNA*	Индивидуальный номер* *Individual number**	Хромосомная локализация** *Chromosomal localization***	Краткое обозначение в исследовании *Brief designation in the study*
hsa-miR-021-5p	MIMAT0000076	chr17: 59841266-59841337	miR_21
hsa-miR-031-5p	MIMAT0000089	chr9: 21512115-21512185	miR_31
hsa-miR-125a-3p	MIMAT0004602	chr19: 51693254-51693339	miR_125a
hsa-miR-138-5p	MIMAT0000430	chr3: 44114212-44114310	miR_138
hsa-miR-144-5p	MIMAT0004600	chr17: 28861533-28861618	miR_144
hsa-miR-146b-5p	MIMAT0002809	chr10: 102436512-102436584	miR_146b
hsa-miR-181b-5p	MIMAT0000257	chr1: 198858873-198858982	miR_181b
hsa-miR-187-3p	MIMAT0000262	chr18: 35904818-35904926	miR_187
hsa-miR-197-3p	MIMAT0000227	chr1: 109598893-109598967	miR_197
hsa-miR-199b-5p	MIMAT0000263	chr9: 128244721-128244830	miR_199b
hsa-miR-200b-3p	MIMAT0000318	chr1: 1167104-1167198	miR_200b
hsa-miR-200a-3p	MIMAT0000682	chr1: 1167863-1167952	miR_200a
hsa-miR-205-5p	MIMAT0000266	chr1: 209432133-209432242	miR_205
hsa-miR-221-3p	MIMAT0000278	chrX: 45746157-45746266	miR_221
hsa-miR-222-3p	MIMAT0000279	chrX: 45747015-45747124	miR_222
hsa-miR-375-3p	MIMAT0000728	chr2: 219001645-219001708	miR_375
hsa-miR-574-3p	MIMAT0003239	chr4: 38868032-38868127	miR_574
hsa-miR-885-5p	MIMAT0004947	chr3: 10394489-10394562	miR_885

Примечание: *https://www.mirbase.org/
**https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/

На полученном наборе данных выполнен поиск пар кандидатов в референсные микроРНК по алгоритмам BestKeeper, NormFinder, geNorm, NormiRAZOR и предложенному алгоритму MDseek, отобраны пары кандидатов в референсные микроРНК (табл. 2). Комбинация значений референсных микроРНК рассчитывалась как арифметическое среднее.

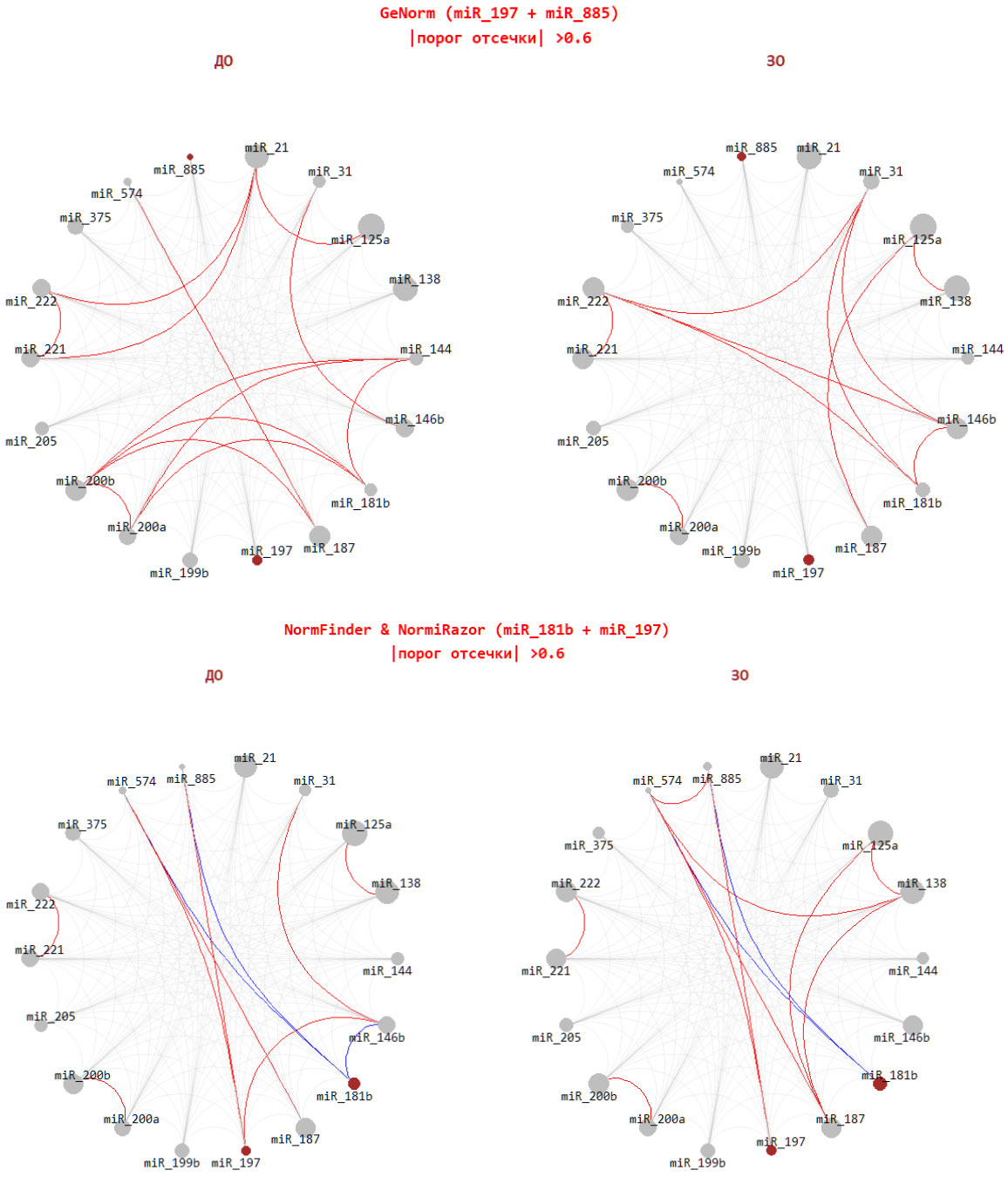
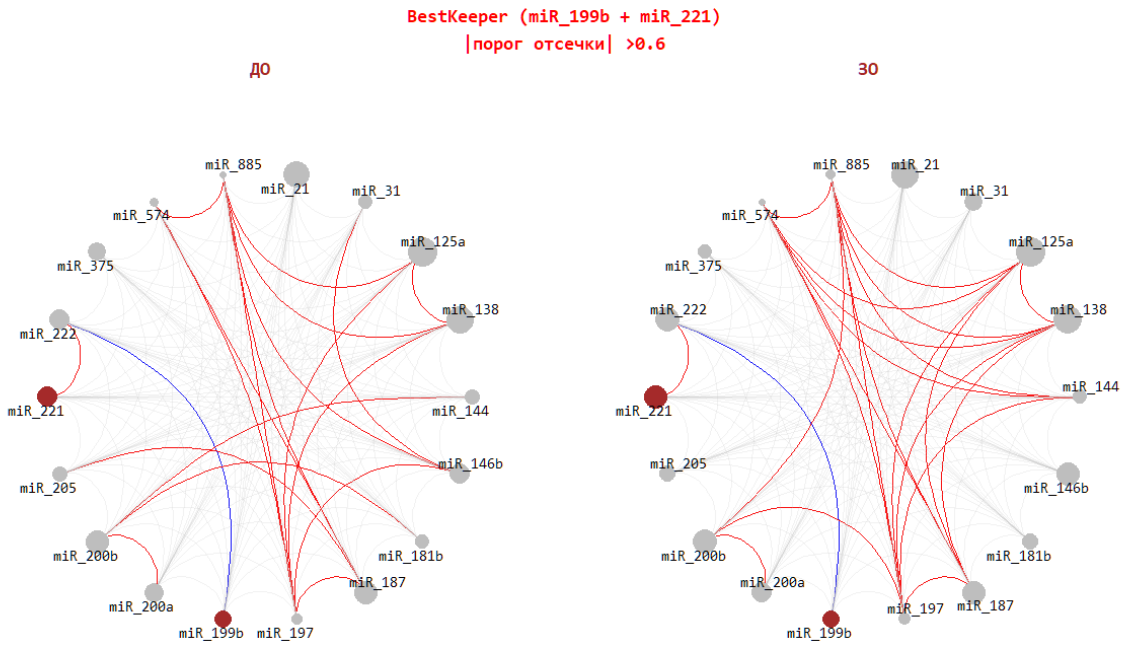
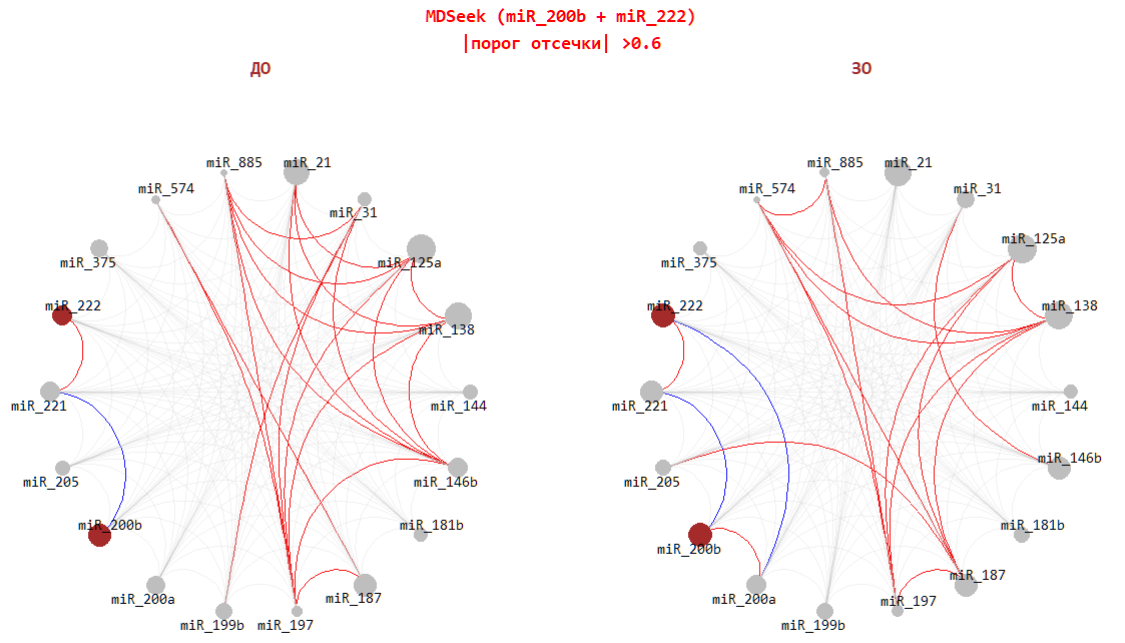


Рис. 1. Корреляционная структура матриц $\|\Delta Ct\|$,

Fig. 1. Correlation structure of matrices $\|\Delta Ct\|$,





полученных по различным алгоритмам

obtained by different algorithms

Для каждого из исследуемых алгоритмов определены пары микроРНК, которые можно считать кандидатами в референсные гены, и рассчитаны матрицы нормализованных значений $\|\Delta Ct\|$ в пределах каждого алгоритма.

Таблица 2
Перечень пар-кандидатов в референсные микроРНК по различным алгоритмам
Table 2
List of the best reference microRNA pairs according to different algorithms

Алгоритм *Algorithm*	Лучшее сочетание пар микроРНК *The best combination of microRNA pairs*
BestKeeper	miR_199b + miR_221
NormFinder	miR_181b + miR_197
geNorm	miR_197 + miR_885
NormiRAZOR	miR_181b + miR_197
MDSeek	miR_200b + miR_222

Далее была рассмотрена корреляционная структура микроРНК в группах доброкачественных и злокачественных образцов тканей на базе нормализованных матриц с порогом отсечки коэффициента корреляции по Спирмену на уровне 0,6. Был проведен анализ главных компонент в пределах каждого алгоритма и выбранной пары (табл. 2) среди микроРНК, не попавших в кандидаты референсных (16 нормализованных микроРНК), и проанализирована доля объясненной дисперсии, построена логистическая бинарная регрессия на первых двух компонентах, которые получены на предыдущем шаге, и оценены основные метрики ее производительности.

Корреляционная структура показана на рис. 1. Вычислялась корреляция по Спирмену между парами признаков. Красным цветом выделены связи при значении $\rho > 0,6$, синим – $\rho < -0,6$ на основании различных алгоритмов. Размер узла определяет силу экспрессии, бордовым цветом выделены узлы референсных микроРНК в пределах изучаемого алгоритма. ДО – доброкачественная опухоль, ЗО – злокачественная опухоль.

Как видно на рис. 1, различие патологических процессов нарушает взаимодействие некоторых микроРНК, что может являться основанием для дальнейших исследований биологических процессов. В наибольшей степени это заметно при использовании референсных микроРНК, отобранных по алгоритму MDseek.

На рис. 2 и 3 показаны результаты анализа методом главных компонент – доля объясненной вариации и диаграмма рассеяния двух первых компонент в отношении двух исследуемых классов. В табл. 3 и 4 приведены метрики производительности бинарных логистических моделей, построенных на двух первых компонентах.

Таблица 3
Доля объясненной вариации после применения различных алгоритмов нормализации
Table 3
Proportion of explained variation after applying different normalization algorithms

Алгоритм *Algorithm*	Доля объясненной вариации *Proportion of explained variation*	
	на двух первых компонентах *in the first two components*	на трех первых компонентах *in the first three components*
geNorm	47,5	58,5
NormiRazor, NormFinder & NormFinder with groups	40,7	53,6
BestKeeper	52,3	63,1
MDseek	59,8	70,1

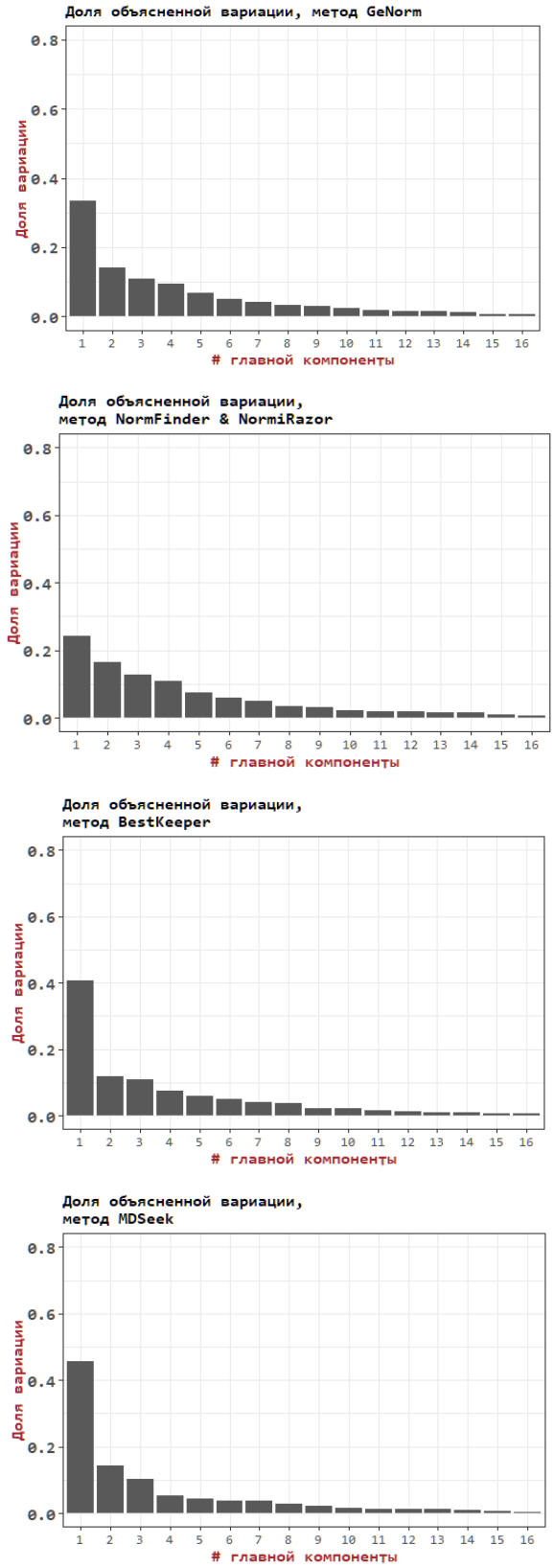


Рис. 2. Вклад компонент в объяснение вариации при использовании различных методов выбора референсных генов

Fig. 2. Component contributions to explaining variation when using different methods for selecting reference genes

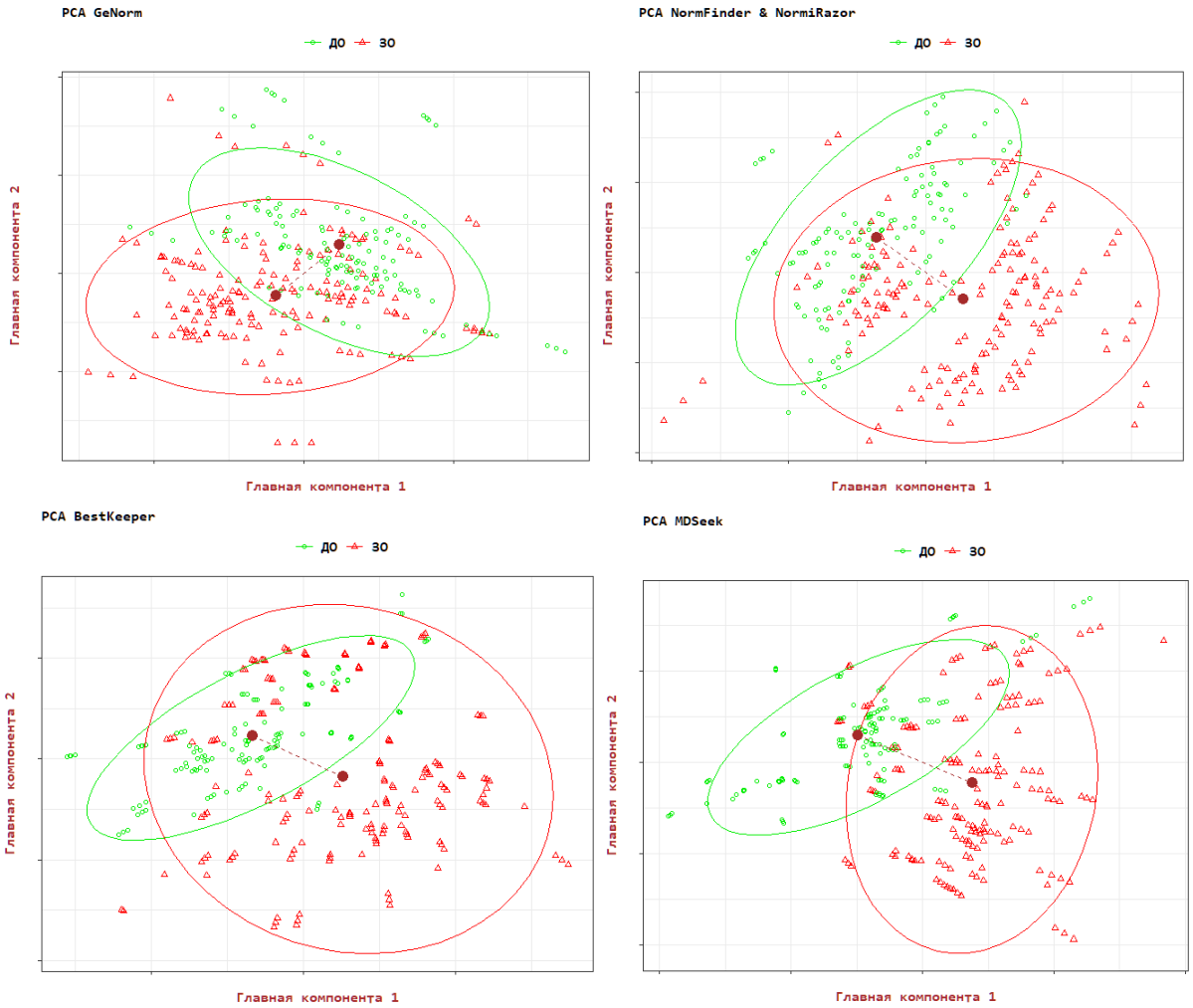


Рис. 3. Диаграмма рассеяния двух первых главных компонент при использовании различных алгоритмов выбора референсных микроРНК

Fig. 3. Scatterplot of the first two principal components for different algorithms for selecting reference microRNA

Таблица 4
Метрики производительности логистической регрессии

Table 4
Logistic regression performance metrics

Алгоритм *Algorithm*	AUC	Чувствительность *Sensitivity*	Специфичность *Specificity*	Прогностическая ценность положительного результата *Positive predictive value*	Прогностическая ценность отрицательного результата *Negative predictive value*	Аккуратность *Accuracy*
geNorm	0,84 (0,80–0,89)	0,77	0,86	0,88	0,74	0,81
NormiRazor, NormFinder & NormFinder with groups	0,90 (0,86–0,93)	0,70	0,95	0,95	0,71	0,81
BestKeeper	0,83 (0,79–0,88)	0,68	0,89	0,68	0,56	0,77
MDseek	0,95 (0,92–0,97)	0,88	0,86	0,89	0,84	0,87

Как следует из табл. 3 и 4, предложенный алгоритм нормализации набора признаков (микроРНК) объясняет больше вариаций и позволяет в дальнейшем строить классификационные модели с лучшими характеристиками производительности.

Полученные результаты позволяют сделать следующие выводы:

1. В отличие от генов микроРНК не являются конститутивными в биологическом организме, что затрудняет или делает невозможным выбор внешних референсных стабильных микроРНК.

2. Корреляция экспрессии отдельных микроРНК с различными биологическими процессами делает возможным учет профиля экспрессии в зависимости от типа ткани, а традиционные алгоритмы нормализации могут быть дополнены компонентами, которые учитывают структуру взаимодействия микроРНК в зависимости от биологического процесса.

3. Предложенный алгоритм MDSeek показал лучшие параметры производительности и может в дальнейшем использоваться для анализа биологических процессов и разработки диагностических биомаркеров (заявка a20250134. Дата приоритета 16.06.2025 г.).

Заключение. Предложен новый алгоритм MDSeek выбора референсных микроРНК при исследовании экспрессии методом кПЦР. Согласно проведенному исследованию данный алгоритм выбора референсных микроРНК для дальнейшей классификации обладает лучшими характеристиками производительности, поскольку учитывает наличие коэкспрессии микроРНК при различных патологических процессах, что может быть использовано для анализа биологических процессов и разработки диагностических биомаркеров. Выбор алгоритма нормализации экспрессии микроРНК критически важен для оптимизации подходов к дифференциальной диагностике опухолей, лечению и прогнозированию течения заболевания, позволяет более эффективно использовать потенциал микроРНК как биомаркеров и терапевтических мишеней в онкологии.

Вклад авторов. *О. В. Красько* – дизайн статьи, разработка алгоритма, сравнение результатов, оформление статьи. *С. В. Якубовский* – участие в получении первичного материала, редактирование текста статьи. *В. Н. Кипень* – проведение молекулярно-генетических исследований, редактирование текста статьи.

References

1. Vandesompele J., De Preter K., Pattyn F., Poppe B., Van Roy N., …, Speleman F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, 2002, vol. 3, pp. 1–12.

2. Karlen Y., McNair A., Perseguers S., Mazza C., Mermod N. Statistical significance of quantitative PCR. *BMC Bioinformatics*, 2007, vol. 8, pp. 1–16.

3. Maltseva D. V., Khaustova N. A., Fedotov N. N., Matveeva E. O., Lebedev A. E., …, Tonevitsky A. G. High-throughput identification of reference genes for research and clinical RT-qPCR analysis of breast cancer samples. *Journal of Clinical Bioinformatics*, 2013, vol. 3, pp. 1–12.

4. Mar J. C., Kimura Y., Schroder K., Irvine K. M., Hayashizaki Y., …, Quackenbush J. Data-driven normalization strategies for high-throughput quantitative RT-PCR. *BMC Bioinformatics*, 2009, vol. 10, pp. 1–10.

5. Bustin S. A., V. Benes, J. A Garson, J. Hellemans, J. Huggett, …, Wittwer C. T. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*, 2009, vol. 55, no. 4, pp. 611–622.

6. Jacob F., Guertler R., Naim S., Nixdorf S., Fedier A., …, Heinzelmann-Schwarz V. Careful selection of reference genes is required for reliable performance of RT-qPCR in human normal and cancer cell lines. *PloS One*, 2013, vol. 8, no. 3, p. e59180.

7. Pfaffl M. W., Tichopad A., Prgomet C., Neuvians T. P. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper – Excel-based tool using pair-wise correlations. *Biotechnology Letters*, 2004, vol. 26, pp. 509–515.

8. Andersen C. L., Jensen J. L., Ørntoft T. F. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Research*, 2004, vol. 64, no. 15, pp. 5245–5250.

9. Grabia S., Smyczynska U., Pagacz K., Fendler W. NormiRazor: tool applying GPU-accelerated computing for determination of internal references in microRNA transcription studies. *BMC Bioinformatics*, 2020, vol. 21, pp. 1–16.

10. Marabita F., de Candia P., Torri A., Tegnér J., Abrignani S., Rossi R. L. Normalization of circulating microRNA expression data obtained by quantitative real-time RT-PCR. *Briefings in Bioinformatics*, 2016, vol. 17, no. 2, pp. 204–212.

11. Friedman R. C., Farh K. K., Burge C. B., Bartel D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 2009, vol. 19, no. 1, pp. 92–105. DOI: 10.1101/gr.082701.108.

12. Iorio M. V., Croce C. M. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Molecular Medicine*, 2012, vol. 4, no. 3, pp. 143–159. DOI: 10.1002/emmm.201100209.

13. Boufraqech M., Klubo-Gwiezdzinska J., Kebebew E. MicroRNAs in the thyroid. *Best Practice & Research Clinical Endocrinology & Metabolism*, 2016, vol. 30, iss. 5, pp. 603–619. DOI: 10.1016/j.beem.2016.10.001.

14. Yoshida K., Yokoi A., Yamamoto Y., Kajiyama H. ChrXq27.3 miRNA cluster functions in cancer development. *Journal of Experimental & Clinical Cancer Research*, 2021, vol. 40, iss. 1, p. 112. DOI: 10.1186/s13046-021-01910-0.

15. Mahalanobis, P. C. On the generalized distance in statistics. *Proceedings of National Institute Science in India*, 1936, vol. 2, pp. 49–55.

16. De Maesschalck R., Jouan-Rimbaud D., Massart D. L. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 2000, vol. 50, no. 1, pp. 1–18.

17. Touloumis A. Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Computational Statistics & Data Analysis*, 2015, vol. 83, pp. 251–261.

Информация об авторах

Красько Ольга Владимировна, кандидат технических наук, доцент, ведущий научный сотрудник, Объединенный институт проблем информатики Национальной академии наук Беларуси.
E-mail: krasko@newman.bas-net.by
https://orcid.org/0000-0002-4150-282X

Якубовский Сергей Владимирович, кандидат медицинских наук, доцент, доцент кафедры хирургии и трансплантологии с курсом повышения квалификации и переподготовки, Белорусский государственный медицинский университет.
E-mail: yakub-2003@yandex.by
https://orcid.org/0000-0003-3759-7050

Кипень Вячеслав Николаевич, кандидат биологических наук, доцент, ведущий научный сотрудник, Институт генетики и цитологии Национальной академии наук Беларуси.
E-mail: v.kipen@igc.by
https://orcid.org/0000-0002-7822-0746

Information about the authors

Olga V. Krasko, Ph. D. (Eng.), Assoc. Prof., Leading Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: krasko@newman.bas-net.by
https://orcid.org/0000-0002-4150-282X

Siarhei U. Yakubouski, Ph. D. (Med.), Assoc. Prof., Assoc. Prof. of Department of Surgery and Transplantology with Advanced Training and Retraining Courses, Belarusian State Medical University.
E-mail: yakub-2003@yandex.by
https://orcid.org/0000-0003-3759-7050

Viachaslau N. Kipen, Ph. D. (Biol.), Assoc. Prof., Leading Researcher, The Institute of Genetics and Cytology of the National Academy of Sciences of Belarus.
E-mail: v.kipen@igc.by
https://orcid.org/0000-0002-7822-0746