

УДК 577.214.5

ОЦЕНКА УРОВНЯ КЛИНИЧЕСКИ ЗНАЧИМЫХ микроРНК НА ОСНОВАНИИ ТОТАЛЬНОГО КЛЕТОЧНОГО ПУЛА РНК АЛГОРИТМАМИ МАШИННОГО ОБУЧЕНИЯ

© 2024 г. Я. В. Соловьев^{1,*}, А. С. Евпак^{1,*}, А. А. Кудряева¹,
академик РАН А. Г. Габибов^{1,3}, А. А. Белогуров^{1,2}

Поступило 28.12.2023 г.

После доработки 25.01.2024 г.

Принято к публикации 03.02.2024 г.

Анализ механизмов возникновения и развития онкологических заболеваний является одной из ключевых задач современной клинической биоинформатики и молекулярной биологии. Омиксные данные пациентов, в частности транскриптом, позволяют довольно подробно описать закономерности в уровне экспрессии и посттранскрипционной регуляции различных типов РНК относительно остального транскриптома. В нашей работе мы создали набор данных, включающий в себя транскриптомные данные около 16 000 пациентов с более чем 160 типами рака, и использовали современные алгоритмы градиентного бустинга, чтобы найти сложные корреляции уровня экспрессии четырех клинически значимых микроРНК, в частности, hsa-mir-21, hsa-let-7a-1, hsa-let-7b и hsa-let-7i, с уровнем экспрессии остальных 60 660 уникальных РНК. Нам удалось показать зависимость уровня экспрессии изучаемых микроРНК от концентрации ряда малых ядерных РНК и регуляторных длинных некодирующих РНК, роль которых в развитии некоторых типов рака была ранее показана экспериментально. Дальнейший анализ созданной базы данных позволит выявить более широкий спектр общих зависимостей изменения уровня экспрессии различных классов РНК при разных онкологических заболеваниях, а также найти уникальные изменения, характерные для отдельных типов злокачественных трансформаций.

Ключевые слова: транскриптом, микроРНК, машинное обучение, градиентный бустинг

DOI: 10.31857/S2686738924030088, **EDN:** VTRGCG

ВВЕДЕНИЕ

МикроРНК имеют длину 18–25 нуклеотидов и представляют собой серию эволюционно консервативных одноцепочечных малых некодирующих молекул РНК [1], которые могут модулировать экспрессию генов как на транскрипционном, так и на посттранскрипционном уровнях посредством связывания с 3'-нетранслируемой областью (3'UTR) целевой матричной РНК (мРНК), что приводит к деградации мРНК или ингибированию трансляции [2]. Будучи негативными регуляторными факторами экспрессии генов, микроРНК могут регулировать онкогенез

и прогрессирование рака, влияя на многочисленные биологические сигнальные пути [3].

Среди многочисленных микроРНК, ассоциированных с развитием рака, микроРНК-21 была одной из первых, которая была идентифицирована как онкогенная микроРНК [4]. Изменение экспрессии микроРНК-21 при раке может быть обусловлено эпигенетическими факторами [5], нарушением транскрипционной и посттранскрипционной регуляции во время биогенеза микроРНК или прямой репрессией микроРНК-21, что приводит к онкогенным фенотипам клеток [6].

В свою очередь, микроРНК-21 участвует в регуляции различных клеточных процессов, играющих важную роль в развитии рака. Так, микроРНК-21 связана с пролиферацией клеток через некоторые из ее мишеней, таких как PDCD4, SPRY2, PTEN и RECK. Было показано, что в клетках рака поджелудочной железы микроРНК-21 может способствовать пролиферации, путем нацеливания на SPRY2, который ингибирует пролиферацию клеток, индуцированную эпидермальным фактором роста (EGF) [7]. Помимо

¹Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова РАН, Москва, Россия

²Российский университет медицины Министерства здравоохранения Российской Федерации, Москва, Россия

³Московский государственный университет им. М.В. Ломоносова, Москва, Россия

*e-mail: solovev@ibch.ru

этого, микроРНК-21 регулирует MAPK/ERK и PI3K/АКТ сигнальные пути, что также оказывает влияние на модуляцию пролиферации клеток [7]. При немелкоклеточном раке легкого также было показано влияние микроРНК-21 на пролиферацию клеток посредством связывания с PDCD4 [8], а при плоскоклеточном раке легких — через PTEN и RECK [9].

Другими значимыми процессами, на которые может влиять микроРНК-21, являются инвазия и метастазирование. Так, в клеточных линиях рака молочной железы, предстательной железы, гепатоцеллюлярной карциномы и рака толстой кишки было обнаружено, что микроРНК-21 оказывает влияние на гены, связанные со снижением метастатического потенциала, такие как PDCD4, TIMP3, TPM1, SERPINB5 и PTEN. Повышение уровня экспрессии микроРНК-21 приводит к подавлению TIMP3, что, в свою очередь, способствует увеличению инвазивности клеточных линий меланомы, но не усиливает их миграционный потенциал [10]. В случае гепатоцеллюлярной карциномы было показано, что микроРНК-21 способна контролировать миграцию и инвазию, связываясь с PTEN, который подавляет экспрессию MMP-9 и MMP-2 посредством дефосфорилирования FAK [11].

Еще одним процессом, регулируемым микроРНК-21, является апоптоз. При раке молочной железы апоптоз модулируется посредством нацеливания микроРНК-21 на BCL2 и PDCD4, что также подтверждается исследованиями *in vitro* с различными типами рака [12]. Было обнаружено, что микроРНК-21 подавляет негативные регуляторы Ras/MAPK-киназы (MEK)/ERK пути, что приводит к ингибированию апоптоза [12].

Было обнаружено, что экспрессия микроРНК-21 повышается при некоторых видах рака, таких как глиобластома, аденокарцинома поджелудочной железы, немелкоклеточный рак легкого и многие другие. В совокупности со способностью микроРНК-21 модулировать значимые для развития рака клеточные процессы, это делает ее важным биомаркером со значительным потенциалом в диагностике и прогнозировании течения рака.

Другим интересным примером участия микроРНК в регуляции процессов, ассоциированных с раком, является семейство микроРНК let-7, которое включает в себя 13 членов, кодирующих 9 зрелых микроРНК [13]. МикроРНК-let-7 относят к категории супрессоров опухолей благодаря способности снижать агрессивность течения рака и резистентность к терапии [14].

Снижение уровня экспрессии микроРНК-let-7 чаще всего связано с плохим прогнозом течения рака, поскольку микроРНК-let-7 участвует в подавлении способности раковых стволовых клеток к проли-

ферации, дифференцировке и самообновлению [15]. Белки семейства Lin28 (Lin28A и Lin28B) могут ингибировать биогенез микроРНК-let-7, связывая предшественника зрелой let-7 (pre-let-7) [14]; также некоторые lncRNA и circRNA могут связывать микроРНК-let-7, ограничивая ее доступ к мРНК генов-мишеней [16].

В поддержании способности к самообновлению и пролиферации раковых стволовых клеток участвует ряд важных сигнальных путей, таких как Wnt/ β -катенин, NOTCH/Hedgehog, STAT3/Nf- κ B, MAPK/ERK и PI3K/АКТ. МикроРНК-let-7 участвует в контроле стволовости рака, действуя как регулятор этих путей посредством петли обратной связи.

Так, например, Wnt/ β -катенин подавляет посттранскрипционную экспрессию let-7 за счет трансактивации Lin28B, супрессора биогенеза микроРНК-let-7 [17]. При этом значительное повышение экспрессии микроРНК-let-7 (например, посредством не прямой активации микроРНК-146a) ингибирует способность к самообновлению раковых стволовых клеток путем подавления передачи сигналов Wnt при раке молочной железы и других видах рака [18].

Аберрантная активация передачи сигналов NOTCH связана с фенотипом стволовости при многих видах рака. Недавние исследования продемонстрировали способность микроРНК семейства let-7 осуществлять регуляцию передачи сигналов в пути NOTCH в нескольких типах опухолевых клеток [19]. Также при раке поджелудочной железы высоко метастатические раковые клетки демонстрируют повышенный уровень экспрессии Shh (sonic hedgehog), который участвует в эмбриональном развитии и поддержании стволовости, что коррелирует с более низкими уровнями экспрессии let-7 [20]. Ингибирование Shh приводит к увеличению уровня микроРНК-let-7 и подавлению экспрессии SOX2, NANOG и EpCAM [21].

Активация факторов транскрипции, таких как STAT3 и Nf- κ B, способствует поддержанию эпигенетической и фенотипической пластичности опухоли и способности к самообновлению за счет подавления микроРНК-let-7 [22, 23]. При этом повышение уровня экспрессии микроРНК-let-7, который действует как вышестоящий супрессор Ras/Nf- κ B сигнального пути, уменьшает злокачественную трансформацию и приобретение свойств раковых стволовых клеток в нормальных кератиноцитах кожи человека [24].

Рецепторные тирозинкиназы, такие как EGFR, активируются внеклеточными лигандами, такими как EGF, с последующей активацией сигнальных путей MAPK/ERK (также известных как Ras-RAF-MEK-ERK) и PI3K-АКТ-mTOR для регуляции пролиферации, самообновления и дифференцировки клеток. Семейство микроРНК let-7 нацелено на H- и K-Ras

для регуляции активности раковых стволовых клеток. Активация экспрессии let-7a приводит к ингибированию K-Ras и подавлению стволового фенотипа раковых клеток [25].

Ряд исследований продемонстрировали способность микроРНК-let-7 и Lin28 модулировать метаболизм глюкозы и жирных кислот. Так, при раке молочной железы, Lin28A и Lin28B усиливают аэробный гликолиз, тогда как let-7 подавляет его посредством регуляции PDK1 HIF-1-независимым образом [26]. Повышенная экспрессия Lin28A/B усиливает синтез жирных кислот *de novo*, способствуя прогрессированию рака через SREBP-1 [27]. Ингибирование Lin28B нарушает рост клеток и метаболизм аминокислот при остром миелолейкозе путем дерепрессии let-7a [28]. Таким образом, подавление экспрессии Lin28 и/или восстановление экспрессии let-7 может обеспечить перспективные терапевтические стратегии лечения рака.

МАТЕРИАЛЫ И МЕТОДЫ

В работе использовался составной набор парных транскриптомных данных для 15964 пациентов с разными онкологическими диагнозами (таблицу с ID номерами пациентов можно получить у автора для переписки), полученный из базы данных Атласа ракового генома: <https://www.cancer.gov/tcga> (англ. *The Cancer Genome Atlas*). Для обучения были отобраны 60660 уникальных РНК, в числе которых были белок-кодирующие, рибосомальные, малые ядерные, малые ядрышковые, длинные некодирующие РНК, а также РНК псевдогенов. Набор микроРНК содержал 1881 уникальный компонент. Для валидации методики на первом этапе нами были отобраны 2 клинически значимые микроРНК: семейство микроРНК-let-7, представленное 3-мя различными членами, и микроРНК-21.

Перед применением алгоритмов статистического анализа и машинного обучения данные были рандомизированы (перемешаны) по строкам для избежания эффекта кластеризации вхождений по типам рака. Для обучения был отобраны 14368 случайных пациентов, валидация моделей осуществлялась на тестовом наборе данных из 1596 пациентов. В обоих датасетах были представлены все имеющиеся типы раков. Для машинного обучения использовалась библиотека градиентного бустинга LightGBM версии 3.3.4 [29]. Подбор оптимальных параметров моделей осуществлялся модифицированным протоколом GridSearchCV из библиотеки sklearn версии 1.2.0 [30]. Модификация состояла в использовании динамической грид-сетки, где величина изменения каждого гиперпараметра зависела от результата работы модели на прошлом этапе. Так, для каждой модели подбирались оптимальные глубина и число

деревьев, а также число “листьев” модели. На первом этапе валидация гиперпараметров осуществлялась на обучающей выборке методом k-fold, данные итеративно разбивались на тренировочные и тестовые подмножества в пропорции 5: 1. Оптимальные модели, в дальнейшем, валидировались на тестовом наборе данных. Для оценки линейной корреляции использовался коэффициент корреляции Пирсона, реализованный в библиотеке NumPy версии 1.23.5 [31]. Для визуализации данных использовалась библиотека Plotly версии 5.6.0 [32].

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

На первом этапе мы оценили качество работы выбранного алгоритма градиентного бустинга для предсказания абсолютных значений уровней экспрессии выбранных микроРНК для тренировочного и тестового набора данных. Оценка включала в себя анализ сонаправленности линий линейной корреляции и значений R^2 между актуальными и предсказанными значениями в обучающей и тестовой выборке. Результаты работы моделей для четырех изучаемых микроРНК представлены на рис. 2 (пункты 1, 3, 5, 7). На основании полученных данных можно утверждать, что после оптимизации гиперпараметров каждой модели алгоритм LightGBM Regressor способен уверенно описать основные зависимости для микроРНК hsa-mir-21 ($R^2 = 0.8$) и для микроРНК hsa-let-7i ($R^2 = 0.86$). Для двух других членов семейства микроРНК hsa-let-7 градиентный бустинг показал чуть менее высокое качество оценки результатов на тестовом наборе данных: $R^2 = 0.63$ для hsa-let-7a-1 и $R^2 = 0.58$ для hsa-let-7b соответственно. Важно отметить, что во всех моделях наблюдается крайне высокая сонаправленность между графиками линейных корреляций для тренировочных и тестовых значений, что говорит об отсутствии значимых не охарактеризованных зависимостей в тренировочных наборах относительно тестового множества. Также следует подчеркнуть важность подбора оптимальных гиперпараметров. Так, в частности, важность оптимизации числа деревьев и числа “листьев”, особенно в случае микроРНК hsa-let-7a-1 и hsa-let-7b, — недостаточное число деревьев быстро приводило к значительному падению качества оценки тестового набора данных. Напротив, избыточное число деревьев решений и значительное число “листьев” несколько ухудшало прогноз за счет переобучения алгоритма, что было особенно заметно на кросс-валидационных тестах, в которых модели одинаковой архитектуры обучались на разных наборах тренировочных и тестовых данных.

Дополнительным подходом оценки качества работы модели машинного обучения является характеристика распределения величины ошибок для тестовых наборов данных, результаты анализа представлены на рис 1.

В нашем случае все модели имеют сходные распределения величин ошибки, только в случае микроРНК hsa-let-7b наблюдается чуть более высокий средний процент ошибок: 44.8% против 37.3% для hsa-mir-21 и 29.6% для hsa-let-7i. В нашем случае, однако, более показательным критерием может выступать величина квантиля 0.9 (максимальная ошибка для 90% выборки, ранжированной по возрастанию величины ошибки) ввиду наличия единичных уникальных артефактных образцов, для которых разница между предсказанными и актуальными значениями составляет до 4 порядков. Так, для микроРНК hsa-mir-21 квантиль 0.9 равен 68.4%, для микроРНК hsa-let-7a-1 – 61.7%, для микроРНК hsa-let-7b – 82% и для микроРНК hsa-let-7i – 63.6%. Отдельно отметим, что попытки кластеризовать и охарактеризовать отдельно имеющийся набор данных по типам рака или по сходству уровня экспрессии РНК для снижения величины ошибки принесли обратный эффект, таким образом, наличие максимально разнообразных данных способствует увеличению качества модели, позволяя выявить закономерности, которые отсутствуют или очень слабо выражены в отдельных типах рака.

Для более детального анализа работы созданных моделей мы выявили и охарактеризовали РНК, имеющие наибольший вклад в качество работы модели. Анализ вклада дескрипторов, ключевых для предсказания членов семейства микроРНК hsa-let-7 и микроРНК hsa-mir-21, представлен на рис. 2 (пункты 2, 4, 6, 8). Основной задачей было оценить биологическую значимость РНК, вносящих основной вклад в работу предсказательной модели. Так, например, для микроРНК-21 количество мРНК тромбоцитарного фактора роста ТУМР является ключевым относительно остальных РНК, ее вклад составляет 25.5%. Клиническая значимость сверхэкспрессии ТУМР ранее была показана для рака почки и глиомы низкой степени злокачественности [33]. Количество мРНК активатора плазминогена урокиназного типа PLAУ и мРНК белка мембраны вакуолей первого типа VMP1, ассоциированного с развитием рака поджелудочной железы [34, 35], также существенно, их вклад составляет 8.25% и 5.5% соответственно. Сверхэкспрессия гена PLAУ ассоциирована с плоскоклеточным раком пищевода [36] и является негативным маркером прогноза времени дожития пациента. Более детальный анализ компонентов модели позволил выявить минорные вклады антисмысловой РНК PDCD4-AS1, мРНК нескольких типов матриксных металло-протеаз (в том числе MMP-2), мРНК ингибитора сериновых эндо-протеаз SERPINB5 и мРНК анти-апоптотического фактора BCL2L12, для которых ранее было показано участие в развитии разных типов онкологии.

Похожий характер распределения наблюдается для микроРНК-let-7i, где ключевую роль играет ма-

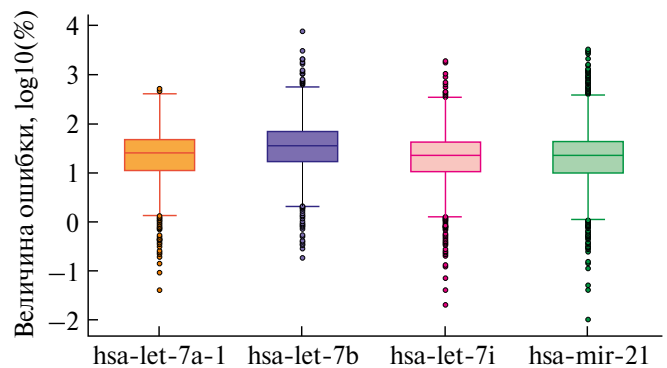


Рис. 1. Распределение величины процента ошибки относительно актуальных значений для тестовых наборов данных.

лая ядрышковая РНК (мяРНК) SNORA49 – ее вклад составляет 52.3%, что на порядок больше следующей по важности РНК белка рибосомы RPL5P8, чья значимость оценивается в 4.5%. Не так давно было показана разница в уровне экспрессии мяРНК SNORA49 в здоровой ткани и при аденокарциноме легкого [37]. Участие рибосомальных белков семейства RPL5 в развитии разных онкологических заболеваний, в том числе рака груди, также было подтверждено недавними исследованиями [38]. При этом ассоциированные с раком яйчников [39] и раком желудка [40] длинные некодирующие РНК LINC01465 и LINC01050 имеют вклад по 0.8% в общее качество модели. Некоторые ранее описанные гены-мишени микроРНК-let-7i также имеют незначительные вклады (менее 0.1%) в модель, в частности, мРНК генов Notch2, PDK1, STAT2 и STAT4, мРНК псевдогенов Nanog 2, 4, 5, 7 и 11, а также мРНК псевдогена Lin28AP1. По-видимому, они являются значимыми для описания небольших групп пациентов, но не являются ключевыми для оценки уровня экспрессии самой микроРНК.

В то же время распределение весов ключевых РНК для микроРНК-let-7a-1 имеет принципиально другой характер. Все РНК имеют незначительный вес, не более 6.1% для малой ядрышковой РНК SNORA12 (ассоциирована с раком легкого [41]). Для оценки количества этой микроРНК также важны количества РНК рибосомального белка RPL5P8 (2.3%), мРНК фосфат-зависимой убиквитин-лигазы FBXO7 семейства F-box (низкий уровень экспрессии или мутации которой связаны с развитием карциномы эндометрия [42]) и мРНК гомеобокс-содержащего гена MEIS1, для которого ранее была показана связь с развитием миелоидной лейкемии [43].

Уровень экспрессии микроРНК-let-7b зависит, главным образом, от двух РНК: мРНК тяжелой цепи ферритина FTH1 и длинной некодирующей РНК

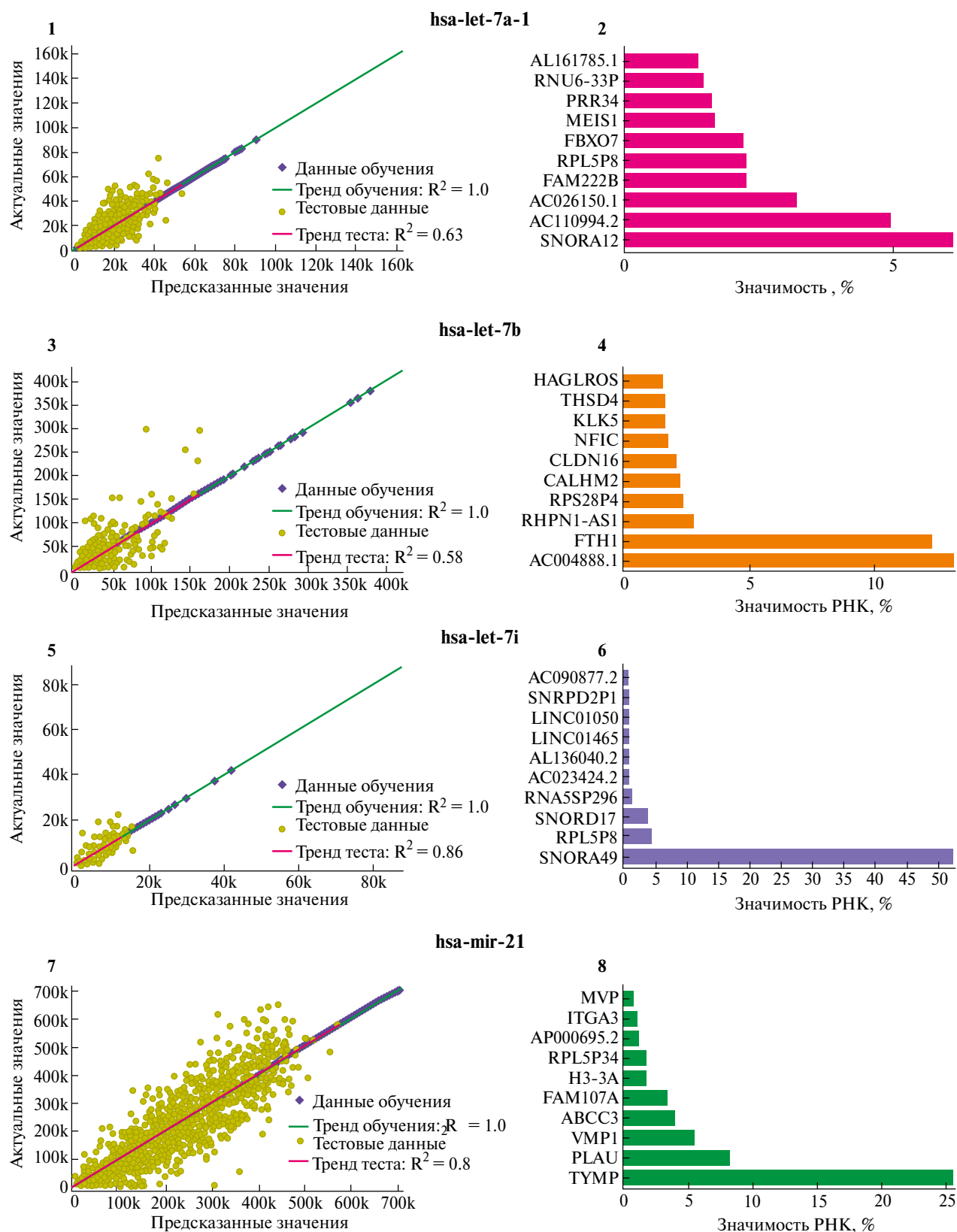


Рис. 2. Анализ качества работы машинного обучения на обучающих и тестовых наборах данных (пункты 1, 3, 5, 7) и анализ вклада ключевых РНК в работу каждой модели (пункты 2, 4, 6, 8).

AC004888.1, вклад каждой из них составляет около 12.5%. Высокий уровень экспрессии FTH1 ассоциирован с благоприятным прогнозом для пациентов с раком груди [44], в тоже время гиперэкспрессия AC004888.1 относительно здоровой ткани наблюдается у пациентов с раком яичников [45]. Следующая по важности антисмысловая РНК RHPN1-AS1, ассоциированная с увеальной меланомой и раком яичников [46, 47], имеет вклад существенно ниже — около 2.75%.

Для сравнения важности отдельных РНК для обучения моделей градиентного бустинга и степенью их линейной корреляции с изучаемой микроРНК мы оценили коэффициенты корреляции Пирсона между четырьмя изучаемыми микроРНК и 60660 остальными РНК. Ранжированные гистограммы приведены на рис. 3.

Для микроРНК hsa-let-7b РНК генов FTH1 и AC004888.1 имеют одни из наибольших коэффициентов корреляции, что совпадает с оценкой их важности алгоритмом градиентного бустинга. То же можно сказать о коэффициенте корреляции РНК гена TYMP для микроРНК hsa-mir-21, а также микроРНК SNORA49 и SNORA12 для микроРНК hsa-let-7i

и микроРНК hsa-let-7a-1 соответственно. Однако нужно особо подчеркнуть, что даже для приведенных выше примеров величина коэффициента Пирсона соответствует слабой корреляции и даже в случае микроРНК SNORA49 и микроРНК hsa-let-7i едва ли соответствует средней. В целом, следует отметить, что за исключением микроРНК hsa-let-7i все остальные микроРНК имеют слабые линейные корреляции с другими РНК, что накладывает существенные ограничения на выбор алгоритмов машинного обучения. Кроме того, очевидно, что простая линейная зависимость не отражает сложных зависимостей, присутствующих в анализируемом датасете — РНК с лучшими коэффициентами корреляции практически не вносят существенного вклада в работу моделей градиентного бустинга. С другой стороны, следует отметить важность минорных компонентов — при попытке удаления из обучающего набора данных тех или иных РНК с следовым вкладом в качестве модели возникали новые артефактные предсказания. Это свидетельствует о том, что минорные компоненты критически необходимы для качественной характеристики небольших групп пациентов, имеющих уникальный паттерн экспрессии некоторых РНК.

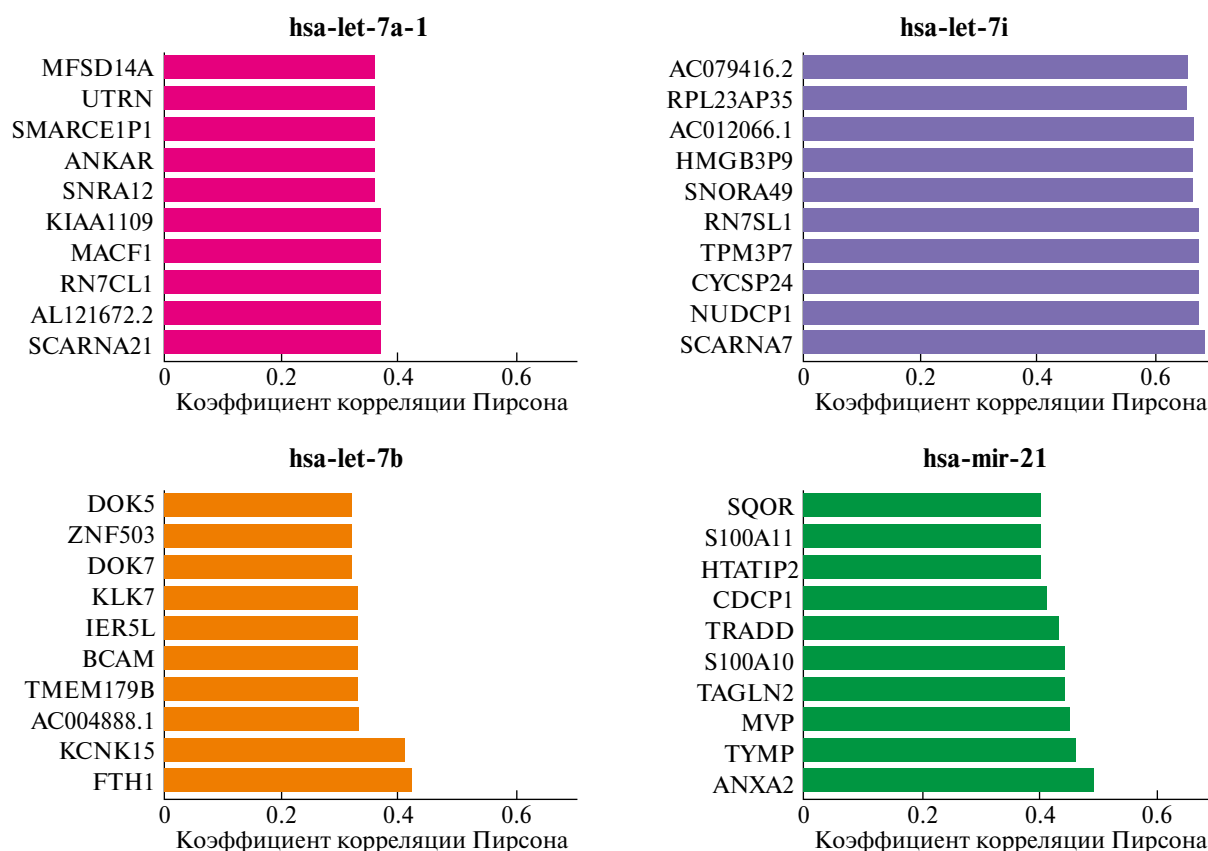


Рис. 3. Соотношение 10 РНК, имеющих наибольшие абсолютные значения коэффициентов корреляции Пирсона, с целевыми микроРНК.

ЗАКЛЮЧЕНИЕ

Оценка абсолютных значений уровней экспрессии клинически значимых микроРНК является важной биоинформатической задачей, имеющей непосредственное практическое применение. Однако не менее важная работа состоит в поиске закономерностей между разными типами РНК, оценке их взаимного влияния в норме и при различных патологиях. В рамках нашего исследования мы показали наличие крайне сложных закономерностей между четырьмя клинически значимыми микроРНК и остальным транскриптомом. Адекватное описание подобных механизмов взаимного влияния является вызовом даже для самых современных алгоритмов машинного обучения. Несмотря на это, градиентный бустинг LightGBM смог выявить основные зависимости для микроРНК hsa-let-7i и микроРНК hsa-mir-21, показав высокую предсказательную силу моделей на тестовых наборах данных. Для микроРНК hsa-let-7b и hsa-let-7a-1 удалось обнаружить только основные зависимости, которые, однако, дают новое представление о механизмах их регуляции при изучаемых типах рака.

Важно отметить, что несмотря на значительную предсказательную силу каждой полученной модели, все они имеют разный паттерн соотношений вкладов ключевых РНК в предсказание выбранной микроРНК. В сочетании с низкими уровнями линейной корреляции Пирсона между ключевыми РНК и целевой микроРНК это позволяет сделать вывод о крайне нетривиальных зависимостях между уровнями экспрессии микроРНК и остальными РНК в клетке. Кроме того, многие важные для моделей градиентного бустинга РНК являются не-белок-кодирующими, что говорит о комплексной системе взаимной регуляции между разными типами некодирующих РНК. Большая часть приведенных выше некодирующих РНК ассоциирована с определенным типом рака, что дает основание предполагать наличие специфических взаимосвязей между регуляторными РНК, уникальных для разных тканей [34, 48, 49]. Вероятно, нарушение уровня специфических не-белок-кодирующих РНК может приводить к развитию онкологии в той же мере, как и более широко изученные отклонения в уровне экспрессии ряда транскрипционных факторов и циклинов. Используемый в нашей работе набор данных включал в себя более двух десятков различных онкологических заболеваний, что позволило выявить на первом этапе ключевые закономерности между уровнями экспрессии изучаемых микроРНК при различных злокачественных новообразованиях. Более детальный анализ вкладов в описание конкретных онкологических заболеваний позволит выявить дополнительные РНК, имеющие минорные вклады в существующей универсальной модели. Кроме того, дальнейшее расширение набо-

ра изучаемых микроРНК позволит выявить больше не-белок-кодирующих РНК, ассоциированных с развитием разных типов рака.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при поддержке гранта РНФ 17-74-30019.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

В данной работе отсутствуют исследования человека или животных.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы данной работы заявляют, что у них нет конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. *Lorenzi L., et al.* The RNA Atlas Expands the Catalog of Human Non-coding RNAs // *Nat. Biotechnol.* 2021. V. 39. № 11. P. 1453–1465. doi:10.1038/s41587-021-00936-1
2. *Jens M., Rajewsky N.* Competition between Target Sites of Regulators Shapes Post-transcriptional Gene Regulation // *Nat. Rev. Genet.* 2015. V. 16. № 2. P. 113–126. doi: 10.1038/nrg3853
3. *Lee Y.S., Dutta A.* MicroRNAs in Cancer // *Annu. Rev. Pathol. Mech. Dis.* 2009. V. 4. № 1. P. 199–227. doi: 10.1146/annurev.pathol.4.110807.092222
4. *Lagos-Quintana M., Rauhut R., Lendeckel W., Tuschl T.* Identification of Novel Genes Coding for Small Expressed RNAs // *Science.* 2001. V. 294. № 5543. P. 853–858. doi: 10.1126/science.1064921
5. *Ferraro A., et al.* Epigenetic Regulation of miR-21 in Colorectal Cancer: ITGB4 as a Novel miR-21 Target and a Three-gene Network (miR-21-ITGB4-PDCD4) as Predictor of Metastatic Tumor Potential // *Epigenetics.* 2014. V. 9. № 1. P. 129–141. doi: 10.4161/epi.26842
6. *Kumarswamy R., Volkmann I., Thum T.* Regulation and Function of miRNA-21 in Health and Disease // *RNA Biol.* 2011. V. 8. № 5. P. 706–713. doi: 10.4161/rna.8.5.16154
7. *Zhao Q., et al.* miR-21 Promotes EGF-induced Pancreatic Cancer Cell Proliferation by Targeting Spry2 // *Cell Death Dis.* 2018. V. 9. № 12. P. 1157. doi: 10.1038/s41419-018-1182-9
8. *Yang Y., et al.* Downregulation of microRNA-21 Expression Restrains Non-small Cell Lung Cancer Cell Proliferation and Migration through Upregulation of Programmed Cell Death 4 // *Cancer Gene Ther.* 2015. V. 22. № 1. P. 23–29. doi: 10.1038/cgt.2014.66
9. *Xu L., Wu Z., Chen Y., Zhu Q., Hamidi S., Navab R.* MicroRNA-21 (miR-21) Regulates Cellular Proliferation, Invasion, Migration, and Apoptosis by Targeting PTEN, RECK and Bcl-2 in Lung Squamous Carcinoma, Gejiu City, China // *PLoS ONE.* 2014. V. 9. № 8. P. e103698. doi: 10.1371/journal.pone.0103698
10. *Martin Del Campo S.E., et al.* MiR-21 Enhances Melanoma Invasiveness via Inhibition of Tissue Inhibitor of Metalloproteinases 3 Expression: In Vivo Effects of MiR-21 Inhibitor // *PLOS ONE.* 2015. V. 10. № 1. P. e0115919. doi: 10.1371/journal.pone.0115919
11. *Meng F., Henson R., Wehbe-Janek H., Ghoshal K., Jacob S. T., Patel T.* MicroRNA-21 Regulates Expression of the PTEN Tumor Suppressor Gene in Human Hepatocellular

- Cancer // Gastroenterology. 2007. V. 133. № 2. P. 647–658. doi: 10.1053/j.gastro.2007.05.022
12. *Hatley M.E., et al.* Modulation of K-Ras-Dependent Lung Tumorigenesis by MicroRNA-21 // *Cancer Cell*. 2010. V. 18. № 3. P. 282–293. doi: 10.1016/j.ccr.2010.08.013
 13. *Lee H., Han S., Kwon C.S., Lee D.* Biogenesis and Regulation of the let-7 miRNAs and Their Functional Implications // *Protein Cell*. 2016. V. 7. № 2. P. 100–113. doi: 10.1007/s13238-015-0212-y
 14. *Balzeau J., Menezes M.R., Cao S., Hagan J.P.* The LIN28/let-7 Pathway in Cancer // *Front. Genet*. 2017. V. 8. doi: 10.3389/fgene.2017.00031
 15. *Yu F., et al.* let-7 Regulates Self Renewal and Tumorigenicity of Breast Cancer Cells // *Cell*. 2007. V. 131. № 6. P. 1109–1123. doi: 10.1016/j.cell.2007.10.054
 16. *Kallen A.N., et al.* The Imprinted H19 LncRNA Antagonizes Let-7 MicroRNAs // *Mol. Cell*. 2013. V. 52. № 1. P. 101–112. doi: 10.1016/j.molcel.2013.08.027
 17. *Cai W.-Y., et al.* Wnt/ β -catenin Pathway Represses let-7 microRNAs Expression via Transactivation of Lin28 to Augment Breast Cancer Stem Cell Expansion // *J. Cell Sci*. 2013. P. jcs.123810. doi: 10.1242/jcs.123810
 18. *Liang R., et al.* MiR-146a Promotes the Asymmetric Division and Inhibits the Self-renewal Ability of Breast Cancer Stem-like Cells via Indirect Upregulation of Let-7 // *Cell Cycle*. 2018. V. 17. № 12. P. 1445–1456. doi: 10.1080/15384101.2018.1489176
 19. *Bao B., et al.* Metformin Inhibits Cell Proliferation, Migration and Invasion by Attenuating CSC Function Mediated by Deregulating miRNAs in Pancreatic Cancer Cells // *Cancer Prev. Res. (Phila. Pa.)*. 2012. V. 5. № 3. P. 355–364. doi: 10.1158/1940-6207.CAPR-11-0299
 20. *Luo G., et al.* Highly Lymphatic Metastatic Pancreatic Cancer Cells Possess Stem Cell-like Properties // *Int. J. Oncol*. 2013. V. 42. № 3. P. 979–984. doi: 10.3892/ijo.2013.1780
 21. *Ahmad A., et al.* Inhibition of Hedgehog Signaling Sensitizes NSCLC Cells to Standard Therapies through Modulation of EMT-regulating miRNAs // *J Hematol Oncol*. 2013. V. 6. № 1. P. 77. doi: 10.1186/1756-8722-6-77
 22. *Alam M., Ahmad R., Rajabi H., Kufe D.* MUC1-C Induces the LIN28B-LET-7-HMGA2 Axis to Regulate Self-Renewal in NSCLC // *Mol. Cancer Res*. 2015. V. 13. № 3. P. 449–460. doi: 10.1158/1541-7786.MCR-14-0363
 23. *Guo L., et al.* Stat3-coordinated Lin-28-let-7-HMGA2 and miR-200-ZEB1 Circuits Initiate and Maintain Oncostatin M-driven Epithelial-Mesenchymal Transition // *Oncogene*. 2013. V. 32. № 45. P. 5272–5282. doi: 10.1038/onc.2012.573
 24. *Jiang R., et al.* The Acquisition of Cancer Stem Cell-like Properties and Neoplastic Transformation of Human Keratinocytes Induced by Arsenite Involves Epigenetic Silencing of let-7c via Ras/NF- κ B // *Toxicol. Lett*. 2014. V. 227. № 2. P. 91–98. doi: 10.1016/j.toxlet.2014.03.020
 25. *Appari M., Babu K.R., Kaczorowski A., Gros W., Her I.* Sulforaphane, Quercetin and Catechins Complement Each Other in Elimination of Advanced Pancreatic Cancer by miR-let-7 Induction and K-ras Inhibition // *Int. J. Oncol*. 2014. V. 45. № 4. P. 1391–1400. doi: 10.3892/ijo.2014.2539
 26. *Ma X., et al.* Lin28/let-7 Axis Regulates Aerobic Glycolysis and Cancer Progression via PDK1 // *Nat. Commun*. 2014. V. 5. № 1. P. 5212. doi: 10.1038/ncomms6212
 27. *Zhang Y., et al.* Lin28 Enhances *De Novo* Fatty Acid Synthesis to Promote Cancer Progression via SREBP-1 // *EMBO Rep*. 2019. V. 20. № 10. P. e48115. doi: 10.15252/embr.201948115
 28. *Zhou J., et al.* Inhibition of LIN28B Impairs Leukemia Cell Growth and Metabolism in Acute Myeloid Leukemia // *J. Hematol. Oncol.* *J Hematol Oncol*. 2017. V. 10. № 1. P. 138. doi: 10.1186/s13045-017-0507-y
 29. *Ke G., et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree // *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*. Red Hook, NY, USA: Curran Associates Inc., 2017. P. 3149–3157.
 30. *Pedregosa F., et al.* Scikit-learn: Machine Learning in Python // *J. Mach. Learn. Res*. 2011. V. 12. P. 2825–2830.
 31. *Harris C.R., et al.* Array Programming with NumPy // *Nature*. 2020. V. 585. № 7825. P. 357–362. doi: 10.1038/s41586-020-2649-2
 32. Plotly Technologies Inc. Collaborative data science. Montréal, QC, 2015. <https://plot.ly>. Ссылка активна на 10 июля 2024 г.
 33. *Yang Y., et al.* A Comprehensive Pan-cancer Analysis on the Immunological Role and Prognostic Value of TYMP in Human Cancers // *Transl. Cancer Res*. 2022. V. 11. № 9. P. 3187–3208. doi: 10.21037/tcr-22-502
 34. *Blum A.E., et al.* RNA Sequencing Identifies Transcriptionally Viable Gene Fusions in Esophageal Adenocarcinomas // *Cancer Res*. 2016. V. 76. № 19. P. 5628–5633. doi: 10.1158/0008-5472.CAN-16-0979
 35. *Vaccaro M.I., Mitchell F., Rivera F., Gonzalez C.D.* Protein Expression in Exocrine Pancreatic Diseases. Focus on VMP1 Mediated Autophagy // *Advances in Protein Chemistry and Structural Biology*. V. 132. Elsevier, 2022. P. 175–197. doi: 10.1016/bs.apcsb.2022.07.001
 36. *Fang L., et al.* PLAU Directs Conversion of Fibroblasts to Inflammatory Cancer-associated Fibroblasts, Promoting Esophageal Squamous Cell Carcinoma Progression via uPAR/Akt/NF- κ B/IL8 Pathway // *Cell Death Discov*. 2021. V. 7. № 1. P. 32. doi: 10.1038/s41420-021-00410-6
 37. *Roberts A.G.K., Catchpole D.R., Kennedy P.J.* Identification of Differentially Distributed Gene Expression and Distinct Sets of Cancer-related Genes Identified by Changes in Mean and Variability // *NAR Genomics Bioinforma*. 2022. V. 4. № 1. P. lqab124. doi: 10.1093/nargab/lqab124
 38. *Fancello L., Kampen K.R., Hofman I.J.F., Verbeek J., Keersmaecker K.D.* The Ribosomal Protein Gene RPL5 is a Haploinsufficient Tumor Suppressor in Multiple Cancer Types // *Oncotarget*. 2017. V. 8. № 9. P. 14462–14478. doi: 10.18632/oncotarget.14895
 39. *Malgundkar S.H., et al.* Identification and Validation of a Novel Long Non-coding RNA (LINC01465) in Ovarian Cancer // *Hum. Cell*. 2022. V. 36. № 2. P. 762–774. doi: 10.1007/s13577-022-00842-x
 40. *Ji Z., et al.* C-Myc-activated Long Non-coding RNA LINC01050 Promotes Gastric Cancer Growth and Metastasis by Sponging miR-7161-3p to Regulate SPZ1 Expression // *J. Exp. Clin. Cancer Res*. 2021. V. 40. № 1. P. 351. doi: 10.1186/s13046-021-02155-7
 41. *Gao L., et al.* Genome-wide Small Nucleolar RNA Expression Analysis of Lung Cancer by Next-generation Deep Sequencing // *Int. J. Cancer*. 2015. V. 136. № 6. doi: 10.1002/ijc.29169
 42. *Zhang H., et al.* FBXO7, a Tumor Suppressor in Endometrial Carcinoma, Suppresses INF2-associated Mitochondrial Division // *Cell Death Dis*. 2023. V. 14. № 6. P. 368. doi: 10.1038/s41419-023-05891-0
 43. *Okada Y., et al.* Homeodomain Proteins MEIS1 and PBXs Regulate the Lineage-specific Transcription of the Platelet Factor 4 Gene // *Blood*. 2003. V. 101. № 12. P. 4748–4756. doi: 10.1182/blood-2002-02-0380

44. Ali A., et al. Ferritin Heavy Chain (FTH1) Exerts Significant Antigrowth Effects in Breast Cancer Cells by Inhibiting the Expression of c-MYC // *FEBS Open Bio*. 2021. V. 11. № 11. P. 3101–3114. doi: 10.1002/2211-5463.13303
45. Meng L., Zhang Q., Huang X. Abnormal 5-methylcytosine lncRNA Methyome is Involved in Human High-grade Serous Ovarian Cancer // *Am. J. Transl. Res.* 2021. V. 13. № 12. P. 13625–13639.
46. Lu L., et al. The Long Non-Coding RNA RHPN1-AS1 Promotes Uveal Melanoma Progression // *Int. J. Mol. Sci.* 2017. V. 18. № 1. P. 226. doi: 10.3390/ijms18010226
47. Wang J., et al. Long Non-coding RNA RHPN1-AS1 Promotes Tumorigenesis and Metastasis of Ovarian Cancer by Acting as a ceRNA against miR-596 and Upregulating LETM1 // *Aging*. 2020. V. 12. № 5. P. 4558–4572. doi: 10.18632/aging.102911
48. Qian Y., Shi L., Luo Z. Long Non-coding RNAs in Cancer: Implications for Diagnosis, Prognosis, and Therapy // *Front. Med.* 2020. V. 7. P. 612393. doi: 10.3389/fmed.2020.612393
49. Chen X., Sun Z. Novel lincRNA Discovery and Tissue-Specific Gene Expression across 30 Normal Human Tissues // *Genes*. 2021. V. 12. № 5. P. 614. doi: 10.3390/genes12050614

EVALUATION OF CLINICALLY SIGNIFICANT miRNAs LEVEL BY MACHINE LEARNING APPROACHES UTILIZING TOTAL TRANSCRIPTOME DATA

**Ya. V. Solov'ev^{a, #}, A. S. Evpak^{a, #}, A. A. Kudriaeva^a,
Academician of the RAS A. G. Gabibov^{a, c}, A. A. Belogurov, Jr.^{a, b}**

^a*Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences,
Moscow, Russian Federation*

^b*Moscow State University of Medicine and Dentistry, Moscow, Russian Federation*

^c*M. V. Lomonosov Moscow State University, Moscow, Russian Federation*

[#]*e-mail: solov'ev@ibch.ru*

Analysis of the mechanisms underlying the occurrence and progression of cancer represents a key objective in contemporary clinical bioinformatics and molecular biology. Utilizing omics data, particularly transcriptomes, enables a detailed characterization of expression patterns and post-transcriptional regulation across various RNA types relative to the entire transcriptome. Here, we assembled a dataset comprising transcriptomic data from approximately 16,000 patients encompassing over 160 types of cancer. We employed state-of-the-art gradient boosting algorithms to discern intricate correlations in the expression levels of four clinically significant microRNAs, specifically hsa-mir-21, hsa-let-7a-1, hsa-let-7b, and hsa-let-7i, with the expression levels of the remaining 60,660 unique RNAs. Our analysis revealed a dependence of the expression levels of the studied microRNAs on the concentrations of several small nucleolar RNAs and regulatory long non-coding RNAs. Notably, the roles of these RNAs in the development of specific cancer types had been previously established through experimental evidence. Subsequent evaluation of the created database will facilitate the identification of a broader spectrum of overarching dependencies related to changes in the expression levels of various RNA classes in diverse cancers. In future it will make possible discovery of unique alterations specific to certain types of malignant transformations.

Keywords: transcriptome, miRNA, machine learning, gradient boosting