

ISSN 1029-8940 (Print)

ISSN 2524-230X (Online)

УДК 577.214.5:616-006.446

<https://doi.org/10.29235/1029-8940-2025-70-2-146-160>

Поступила в редакцию 09.12.2024

Received 09.12.2024

Е. В. Гузова, А. А. Гавричков, И. Н. Ильюшёнков, Т. В. Романовская,
М. А. Лиховец, В. В. Скакун, Н. Н. Яцков, В. В. Гринев

Белорусский государственный университет, Минск, Республика Беларусь

ВЛИЯНИЕ СОХРАНЯЕМЫХ ИНТРОНОВ НА КОДИРУЮЩИЙ ПОТЕНЦИАЛ МОЛЕКУЛ РНК В ЛЕЙКОЗНЫХ КЛЕТКАХ ЧЕЛОВЕКА

Аннотация. Сохранение интронов – один из типов альтернативного сплайсинга, который подразумевает образование молекул зрелых РНК, содержащих в своей структуре хотя бы одну интронную последовательность. Если интронная последовательность попадает в кодирующую часть мРНК, то она может поменять структуру, свойства и порой функциональные возможности соответствующего белка. Более того, некоторые белки такого рода могут быть иммуногенными и рассматриваются как потенциальные мишени для терапии ряда заболеваний. Однако в целом влияние сохранения интронов на кодирующий потенциал РНК до сих пор изучено слабо.

В настоящей работе изложены результаты комбинированного транскриптомно-протеомного исследования 3540 последовательностей интронов, сохраняемых в зрелых молекулах РНК генов, активных в лейкозных клетках человека. Проведена классификация таких интронов в соответствии с наличием у них кодирующего потенциала и описано 7 возможных способов влияния сохранения интронов на последующую трансляцию соответствующих транскриптов. Выявлено 56 интронных последовательностей, способных транслироваться, причем 3 из них – с образованием ранее неизвестных полипептидов.

Ключевые слова: альтернативный сплайсинг РНК, сохранение интрона, транскриптом, кодирующий потенциал, протеом, биоинформатический анализ

Для цитирования: Влияние сохраняемых интронов на кодирующий потенциал молекул РНК в лейкозных клетках человека / Е. В. Гузова, А. А. Гавричков, И. Н. Ильюшёнков [и др.] // Весті Нацыянальнай акадэміі навук Беларусі. Сэрыя біялагічных навук. – 2025. – Т. 70, № 2. – С. 146–160. <https://doi.org/10.29235/1029-8940-2025-70-2-146-160>

Katsiaryna V. Huzava, Aleksey A. Gavrichkov, Ilya M. Ilyushonak, Tatsiana V. Ramanouskaya,
Maksim A. Likhavets, Victor V. Skakun, Mikalai M. Yatskou, Vasily V. Grinev

Belarusian State University, Minsk, Republic of Belarus

EFFECT OF RETAINED INTRONS ON THE CODING POTENTIAL OF RNA MOLECULES IN HUMAN LEUKEMIA CELLS

Abstract. Intron retention is a mode of alternative splicing in which at least one intron is retained in mature RNA molecules. The inclusion of an intron sequence within the coding region of the mRNA can change the structure, properties and, in some cases, the functions of the corresponding protein. Furthermore, some of these altered proteins can be immunogenic and are considered as potential targets for the treatment of a number of diseases. In general, however, the effect of intron retention on the coding potential of RNA molecules remains poorly studied.

This paper presents the results of a combined transcriptomic-proteomic study of 3540 intron sequences retained in mature RNA molecules of genes that are active in human leukemia cells. The introns were classified according to their coding potential and seven possible ways in which intron retention can influence the subsequent translation of the corresponding transcripts were described. A total of 56 translation-capable intronic sequences were identified, and 3 of them produce previously unknown polypeptides.

Keywords: alternative RNA splicing, intron retention, transcriptome, coding potential, proteome, bioinformatics analysis

For citation: Huzava K. V., Gavrichkov A. A., Ilyushonak I. M., Ramanouskaya T. V., Likhavets M. A., Skakun V. V., Yatskou M. M., Grinev V. V. Effect of retained introns on the coding potential of RNA molecules in human leukemia cells. *Vesti Natsyonal'noi akademii nauk Belarusi. Seriya biyalogichnykh nauk = Proceedings of the National Academy of Sciences of Belarus. Biological series*, 2025, vol. 70, no. 2, pp. 146–160 (in Russian). <https://doi.org/10.29235/1029-8940-2025-70-2-146-160>

Введение. Согласно современным представлениям геном человека насчитывает порядка 19,5 тыс. белок-кодирующих генов (GENCODE, вып. 47, окт. 2024 г.). Каталог конечных продуктов их экспрессии продолжает расширяться, однако уже сейчас в публичных базах данных депонировано свыше 100 тыс. уникальных белковых последовательностей человека, что значительно

превышает упомянутое количество кодирующих их генов [1]. Возможность существования подобного разнообразия белков главным образом обусловлена мозаичностью структуры генов человека. В процессе созревания продукты их транскрипции способны подвергаться альтернативному сплайсингу (АС) – сложному многоступенчатому молекулярному процессу, обеспечивающему образование нескольких изоформ РНК из одного предшественника [2].

Среди известных типов АС наименее изученным и долгое время ошибочно рассматриваемым только в контексте аберрантного функционирования системы сплайсинга РНК является сохранение интронов. При данном типе сплайсинга образующиеся молекулы зрелых РНК в своем составе несут последовательности интронов, называемые сохраненными интронами (от англ. *retained introns*, RI) [3]. С появлением секвенаторов нового поколения, а также с развитием биоинформатических подходов, позволяющих осуществлять анализ больших транскриптомных данных, границы нашего понимания феномена сохранения интронов существенно расширились. К настоящему времени удалось установить, что сохранение интронов в транскриптах – это не только следствие возникающих на молекулярном уровне нарушений, присущих различным заболеваниям (в том числе и онкологическим), но и нормальный процесс, характеризующийся тканевой и клеточной специфичностью. Эти открытия, в свою очередь, обусловили повышение интереса к изучению судьбы транскриптов, образующихся в результате протекания АС указанного типа [4].

Исходно предполагалось, что молекулы зрелых мРНК, содержащие интрон(-ы), не обладают функциональной активностью и элиминируются. Последнее объяснялось содержанием в пределах их последовательностей преждевременных терминирующих кодонов (от англ. *premature termination codons*, PTC), распознающихся компонентами системы нонсенс-опосредованной деградации (от англ. *nonsense-mediated decay*, NMD) мРНК. Однако в дальнейшем выяснилось, что мРНК, содержащие PTC, способны избегать деградации и сохраняться в клетках, внося очевидный вклад в разнообразие транскриптома [5].

Более того, такие молекулы мРНК, несмотря на присутствие интронных последовательностей и PTC, могут транслироваться. В доступных на сегодняшний день научных работах, посвященных изучению феномена сохранения интронов, изложены некоторые факты, касающиеся особенностей трансляции подобных транскриптов и структуры синтезирующихся на их матрице белков. Так, известно, что в зависимости от протяженности интронов, их положения относительно стартового кодона, а также наличия в них стоп-кодонов результатом трансляции несущих их молекул РНК могут быть усеченные или же, наоборот, удлиненные изоформы белков с измененной функцией и/или клеточной локализацией. Более того, отмечается, что в случаях, когда речь идет о протекании патологических процессов, среди подобных белков обнаруживаются и иммуногенные молекулы. Однако все эти данные были получены главным образом при изучении вклада в разнообразие протеома человека тех интронных последовательностей, которые, сохраняясь в молекулах зрелых РНК, локализуются в области главных открытых рамок считывания мРНК (от англ. *main open reading frames*, mORF) и могут быть транслированы [6, 7]. При этом работы, которые были бы сфокусированы на системной оценке влияния всех RI (как транслируемых, так и нетранслируемых) на структуру несущих их транскриптов и соответствующих белков, в настоящее время фактически отсутствуют.

В контексте вышеописанного целью настоящего исследования стала оценка влияния интронов, сохраняемых в зрелых РНК белковых генов, активных в лейкозных клетках человека, на кодирующий потенциал таких молекул мРНК.

Материалы и методы исследования. В основе настоящего исследования лежал анализ прошедших стандартную процедуру предпроцессинга оригинальных, а также публичных данных секвенирования по методу парноконцевого RNA-Seq транскриптомов лейкозных и нормальных клеток крови человека. К группе оригинальных данных, ранее описанных в работе В. В. Гринева с соавт. [8], относились библиотеки, полученные в результате шести независимых повторов секвенирования полного транскриптома клеток модельной линии острого миелоидного лейкоза (ОМЛ) Kasumi-1, в частности клеток Kasumi-1, предварительно трансфицированных короткими интерферирующими РНК (киРНК), обеспечивающими нокдаун гибридного онкогена *RUNX1::RUNX1T1*,

а также контрольных клеток, трансфекция которых проводилась с использованием неактивных киРНК (по три повтора для каждого варианта). Второй набор данных состоял из отобранных среди представленных в базе данных European Nucleotide Archive [9] результатов секвенирования с использованием секвенаторов HiSeq 2000/2500 (Illumina) транскриптомов клеток ОМЛ с транслокацией t(8;21)(q22;q22), с инверсией inv(16), с нормальным кариотипом и перестройками гена *MLL*, а также транскриптомов нормальных CD34⁺ гемопоэтических стволовых/прогениторных клеток из красного костного мозга, периферической и пуповинной крови здоровых доноров (по 20 образцов для каждого типа клеток).

Идентификация геномных координат RI проводилась с использованием оригинальной методики, подробное описание которой приведено в наших более ранних работах [8, 10]. Восстановление нуклеотидных последовательностей обнаруженных RI осуществлялось с использованием функционала R/Bioconductor пакета Biostrings [11]. Их последующая трансляция *in silico* проводилась в трех рамках считывания (+0, +1 и +2) с использованием оригинальной (пользовательской) R-функции translate().

Реконструкция транскриптома клеток Kasumi-1 осуществлялась с помощью пакета StringTie [12]. Предварительно прочтения, полученные в ходе каждого повтора секвенирования РНК, картировались относительно эталонной сборки GRCh38.p7 генома человека с использованием алгоритма subunc из R-пакета Rsubread [13]. Полученные файлы формата BAM совместно с эталонными аннотациями, содержащимися в 85 вып. базы данных Ensembl, использовались в качестве входных данных непосредственно для сборщика. Фильтрация реконструированных полноразмерных РНК выполнялась с помощью оригинальной R-функции filtrateTrans(). Исходный код этой функции, а также другие R-коды, специально разработанные для целей настоящего исследования, находятся в свободном доступе в репозитории GitHub [14].

Для подтверждения экспрессии аминокислотных последовательностей, предсказываемых биоинформатически на основе транскриптомных данных, использовалась нейронная сеть DIA-NN [15]. В качестве протеомных данных выступили результаты масс-спектрометрического анализа (МС-анализа) трипсинолизованной белковой фракции клеток Kasumi-1 (с нокдауном гибридного онкогена *RUNX1::RUNX1T1* или без нокдауна, по 5 повторов на каждый вариант). Протеомные данные были получены на базе Института биологических наук при Ньюкаслском университете (г. Ньюкасл, Великобритания) с использованием масс-спектрометра Orbitrap Fusion Lumos Tribrid. Наряду с данными МС-анализа на вход нейронной сети DIA-NN также подавался файл FASTA, содержащий как результаты трансляции *in silico* RI, так и 105 681 последовательность известных на сегодняшний день белков человека, депонированных в базах данных UniProtKB/Swiss-Prot [16], Consensus CDS [17] и/или NCBI RefSeq [18].

Анализ полученных результатов осуществлялся в среде программирования R с помощью как оригинальных кодов собственной разработки, так и функционала пакетов, депонированных в хранилищах CRAN [19] либо Bioconductor [20].

Результаты и их обсуждение. Первый этап настоящего исследования был нацелен на идентификацию RI протяженностью не менее 50 нуклеотидов, присутствие которых в транскриптоме клеток Kasumi-1 подтверждается данными RNA-Seq. С помощью ранее разработанного нами подхода [8, 10] удалось идентифицировать 3540 RI, принадлежащих 2401 гену. Характер их распределения по классам генетических элементов, аннотированных в базе данных Ensembl, отражен на рис. 1, а.

Среди всех обнаруженных RI 1912 составили группу уже аннотированных в базе данных Ensembl сохраняющихся интронов. Причем 998 из них были ассоциированы с последним экзонном соответствующего гена, 498 – с первым экзоном, 387 – с внутренними экзонами и, наконец, 29 – с одноэкзонными транскриптами. Оставшиеся 1628 RI, согласно генным аннотациям, соответствовали каноническим интронам.

Идентификация RI, обладающих кодирующим потенциалом. Очевидно, что оценка кодирующего потенциала RI может иметь ряд особенностей. Во-первых, как это было подробно описано в наших более ранних работах [8, 10], зачастую RI представляют собой лишь уникальную нишу с чем не перекрывающуюся часть от всей последовательности интрона. Во-вторых, мы заранее

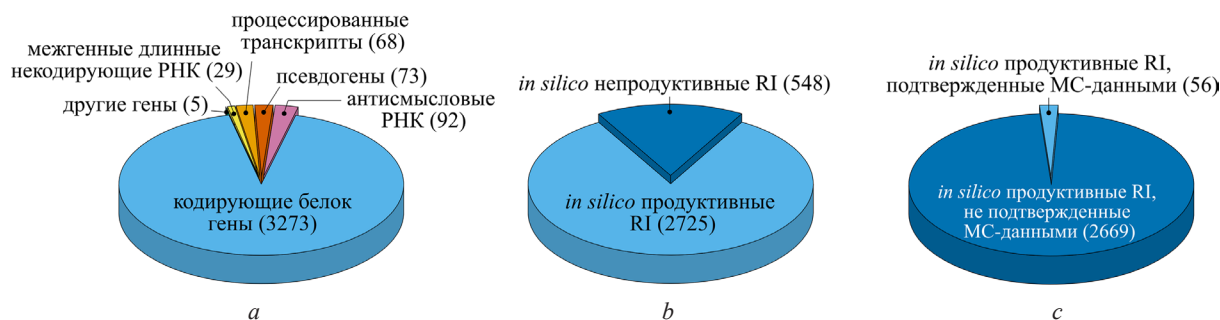


Рис. 1. Белок-кодирующий потенциал RI: *a* – принадлежность RI различным классам генетических элементов; *b* – результаты *in silico* оценки кодирующего потенциала RI, принадлежащих только генам, кодирующим белки; *c* – подтверждаемость МС-данными наличия кодирующего потенциала у *in silico* продуктивных RI

Fig. 1. Protein coding potential of RI: *a* – pie chart showing the assignment of RI to different classes of genetic elements; *b* – *in silico* assessment of the amino acid coding potential of RI from protein-coding genes only; *c* – MS data support the amino acid coding potential of only a minor fraction of RI

не знаем, какова структура транскрипта, несущего в себе тот или иной сохраненный интрон, и каково точное положение последнего в этом транскрипте. Наконец, работая только с RI, невозможно точно определить их локализацию в мРНК относительно стартового кодона. В связи с этим *in silico* трансляция нуклеотидных последовательностей RI проводилась в трех рамках считывания с помощью оригинальной R-функции `translate()`. Наряду с самой трансляцией эта функция позволила осуществить еще и фильтрацию получаемых аминокислотных последовательностей по длине. В настоящем исследовании для успешного прохождения отбора аминокислотные последовательности должны были иметь протяженность 20 аминокислотных остатков (а. о.) и более.

В результате первичной фильтрации из дальнейшего анализа было исключено 603 RI, 548 из которых принадлежали генам, кодирующим белки (рис. 1, *b*). Среди оставшихся 2937 RI 1558 могли быть транслированы *in silico* в более чем одной рамке считывания с образованием продукта длиной, равной или превышающей пороговое значение 20 а. о. В общей сложности на основе таких RI было получено 4705 аминокислотных последовательностей. В дальнейшем мы будем называть их экспериментальными полипептидами в том смысле, что они предсказаны биоинформатически на основе изучения RI, что их длина может быть как меньше, так и больше 50 а. о. и что осуществляется их экспериментальный поиск и обнаружение. Выявление среди них действительно экспрессирующихся в клетках Kasumi-1 проводилось с помощью нейронной сети DIA-NN. Она принимает на вход объединенный файл FASTA и осуществляет гидролиз *in silico* содержащихся в нем аминокислотных последовательностей по сайтам расщепления трипсином, а затем генерирует *in silico* библиотеку масс-спектров полученных пептидов длиной от 7 до 30 а. о. В последующем она используется для идентификации в МС-данных тех пептидов, которые могли бы быть получены в случае экспрессии в исследуемых клетках экспериментальных полипептидов.

Возвращаемые DIA-NN результаты обрабатывались с помощью оригинального R-кода `parseDIANNoutput()`, обеспечивающего фильтрацию детектированных МС-пептидов, спектры которых присутствуют в протеомных МС-данных, и экспериментальных полипептидов по нескольким параметрам. В настоящем исследовании достоверно идентифицированными считались те МС-пептиды, для которых такие показатели, как *q*-значение (*p*-значение, скорректированное по методу Беньямини – Хохберга), а также значение апостериорной вероятности случайного обнаружения оказывались не более 0,01. Кроме того, уровни их интенсивности должны были превышать нулевые значения как минимум в трех из пяти образцов одной группы (с нокдауном гибридного онкогена *RUNX1::RUNX1T1* либо без него). Фильтрация выявленных в клетках Kasumi-1 экспериментальных полипептидов (т. е. тех, чья экспрессия подтверждалась обнаруженными МС-пептидами) осуществлялась также по рассчитанным для них *q*-значениям с пороговым уровнем 0,01. В случае экспериментальных полипептидов, для которых не обнаруживались прототи-

пические МС-пептиды, рассматриваемому показателю автоматически присваивалась 1, и в дальнейшем анализ они не включались. Наконец, на завершающем этапе проводился отбор экспериментальных полипептидов в соответствии с количеством подтверждающих их МС-пептидов. Согласно литературным источникам на один белок должно приходиться по крайней мере два уникальных МС-пептида, чтобы считать его экспрессирующимся в изучаемом типе клеток [21, 22]. Однако в настоящей работе мы снизили этот порог до одного, так как изучаемые экспериментальные полипептиды, вероятнее всего, представляют собой лишь фрагменты полноразмерных белков. В частности, это заметно по распределению их длин (рис. 2, а), значительно смещенному влево в сравнении с распределением длин известных на сегодняшний день белков человека.

В целом из всех RI, обнаруженных нами в транскриптоме клеток Kasumi-1, кодирующими оказались лишь 56 (рис. 1, с), причем все они уже были описаны ранее в базе данных Ensembl именно как RI, но без указаний на их кодирующий потенциал. А вот новые (не аннотированные ранее) RI, также идентифицированные нами в транскриптоме клеток Kasumi-1, не являются кодирующими. То, что большинство RI оказались не кодирующими, по-видимому, может иметь несколько объяснений. Одно из объяснений, на наш взгляд, является тривиальным – недостаточная чувствительность использованного метода протеомного анализа. Так, те RI, для которых мы не обнаружили экспериментальных полипептидов, экспрессируются статистически значимо ниже ($p_{\text{Mann-Whitney}} = 5,0 \times 10^{-12}$), чем кодирующие RI (рис. 2, б). Такую же закономерность мы увидели и в случае с не кодирующими и кодирующими экзонами тех же генов, у которых были обнаружены RI ($p_{\text{Mann-Whitney}} < 2,2 \times 10^{-16}$) (рис. 2, б). При этом пусть и не всегда, но все же прослеживается взаимосвязь между уровнями экспрессии мРНК и соответствующего белка. Так что как минимум часть из тех RI, которых мы отнесли к не кодирующим, может оказаться в конечном итоге кодирующими, если повысится чувствительность методов массовой детекции белков.

Ожидается, что идентифицированные нами кодирующие RI принадлежат мультиэкзонным белковым генам и ассоциированы с первыми, последними или внутренними экзонами транскриптов (11, 33 и 12 RI из 56 соответственно). Интересно отметить, что онтологический анализ [23] полного списка таких генов выявил статистически значимые обогащения по генам, которые кодируют белки ядра и ядерных телец, в том числе ответственных за сплайсинг РНК (см. таблицу).

Идентификация RI, кодирующих новые экспериментальные полипептиды. На данном этапе исследования нерешенным оставался вопрос относительно того, принадлежат обнаруженные нами экспериментальные полипептиды к уже известным белкам или же являются неизвестными белковыми продуктами генов. Чтобы ответить на этот вопрос, мы разработали многоэтапный

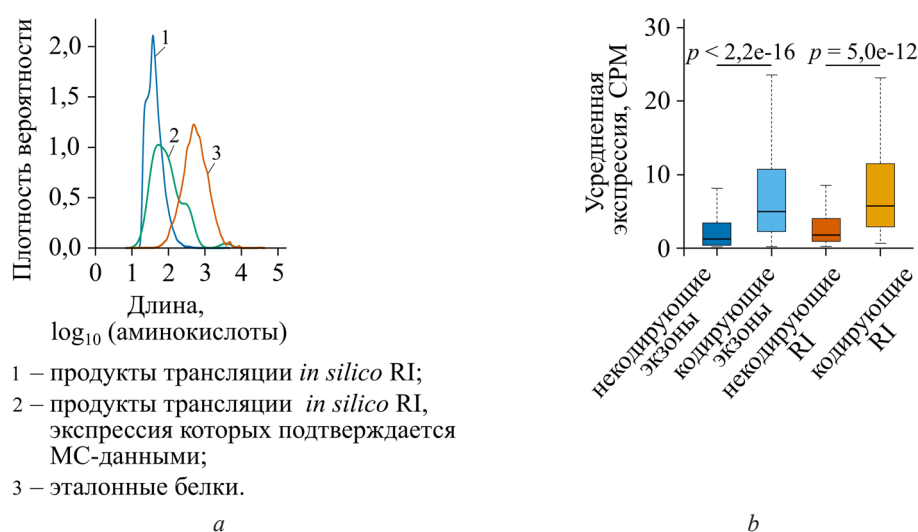


Рис. 2. Распределение длин экспериментальных полипептидов, кодируемых RI (а), а также уровни экспрессии не кодирующих и кодирующих RI (б), обнаруженных в транскриптоме клеток Kasumi-1

Fig. 2. Length distribution of experimental polypeptides encoded by RI (a) and abundance of non-coding as well as coding RI in the transcriptome of Kasumi-1 cells (b)

**Классификация генов, в зрелых молекулах РНК которых были обнаружены кодирующие RI,
в соответствии с генными онтологиями**

Gene ontology classification of genes harboring protein-coding RI

Термин геной онтологіі	Кратность обогащения	Скорректированное <i>p</i> -значение
Рибонуклеопротеиновая гранула (ribonucleoprotein granule)	8,28	1,66E-02
Ядерный спекл (nuclear speckle)	6,33	2,02E-02
Ядерное тельце (nuclear body)	4,49	1,38E-02
Нуклеоплазма (nucleoplasm)	2,62	1,63E-04
Перинуклеарное пространство (nuclear lumen)	2,40	5,55E-04
Просвет внутриклеточной органеллы (intracellular organelle lumen)	1,98	8,46E-03
Просвет органеллы (organelle lumen)	1,98	1,02E-02
Межмембранное пространство (membrane-enclosed lumen)	1,98	7,25E-03
Ядро (nucleus)	1,86	1,24E-03
Внутриклеточная мембранная органелла (intracellular membrane-bounded organelle)	1,47	3,49E-03
Внутриклеточная органелла (intracellular organelle)	1,38	1,38E-02
Мембранная органелла (membrane-bounded organelle)	1,35	3,24E-02

биоинформатический конвейер, исходно требующий наличия двух наборов данных: сиквенсов экспериментальных полипептидов и аминокислотных последовательностей эталонных белков человека. В настоящей работе в качестве последних выступили белки, депонированные в таких базах данных, как UniProtKB/Swiss-Prot, Consensus CDS и NCBI RefSeq.

Последовательности каждого набора разбивались на *k*-меры длиной 8 а. о., и проводилось вычитание *k*-меров эталонных белков из *k*-меров экспериментальных полипептидов. Описанная процедура позволила отобрать только те *k*-меры экспериментальных полипептидов, которые были специфичными для лейкозных клеток. Далее такие *k*-меры собирались обратно в континги (непрерывные фрагменты экспериментальных полипептидов, специфичные для лейкозных клеток). Идентифицированные таким образом специфические фрагменты полипептидов (а также МС-пептиды, подтверждающие их экспрессию в клетках Kasumi-1) дополнительно выравнивались относительно всех известных на сегодняшний момент белков человека с помощью blastp [24]. Данная процедура позволила нам разделить 56 экспериментальных полипептидов, кодируемых RI, на два класса:

1. Принадлежащие эталонным белкам либо идентичные аминокислотным последовательностям неохарактеризованных или плохо охарактеризованных белков из базы данных blastp.
2. Имеющие идентичность менее 100 % по отношению к любым известным белкам человека и, таким образом, являющиеся новыми.

В итоге полипептидов первого класса оказалось 53, второго – 3, кодируемых генами *UBE2M*, *OGT* и *SLC10A3*. Продуктом экспрессии гена *UBE2M* является E2-убиквитин-конъюгирующий белок, способный вносить свой вклад в поддержание клеточного гомеостаза путем обеспечения нормального протекания недиляции – одного из типов пост-трансляционной модификации белков [25]. Ген *OGT* кодирует гликозилтрансферазу, которая осуществляет перенос N-ацетилглюкозамина от уридиндифосфат-N-ацетилглюкозамина к остаткам серина/треонина белков-мишеней, обеспечивая тем самым регуляцию ряда биологических процессов [26]. Ген *SLC10A3* входит в семейство генов, кодирующих белки-транспортеры [27].

Очевидно, что идентификация новых полипептидов – это наиболее значимая часть настоящей работы. Вместе с тем не исключено, что выявленные полипептиды являются частью протеома лишь клеток Kasumi-1 (как артефакт этой клеточной линии), но не клеток ОМЛ вообще. К сожалению, протеомные данные надлежащего качества по первичным клеткам ОМЛ, которые позволили бы напрямую провести поиск детектированных новых пептидов в клетках пациентов, отсутствуют в публичных базах данных. Поэтому мы пошли другим путем и воспользовались

транскриптомными данными для первичных лейкозных и нормальных клеток крови человека как источником косвенной информации о распространенности обнаруженного феномена.

Мы провели расширенный анализ опубликованного нами ранее набора данных транскриптомного секвенирования первичных клеток ОМЛ с транслокацией $t(8;21)(q22;q22)$, с инверсией $inv(16)$, с нормальным кариотипом и перестройками гена *MLL*, а также нормальных $CD34^+$ клеток из красного костного мозга, периферической и пуповинной крови здоровых доноров (по 20 образцов для каждого типа клеток) [8]. Предварительно прошедшие стандартную процедуру предпроцессинга прочтения картировались с помощью алгоритма *subjunc* относительно референсной сборки GRCh38.p7 генома человека. Полученные на выходе BED-файлы содержали информацию обо всех идентифицированных экзон-экзонных стыках (от англ. *exon-exon junctions*, EEJ) и их покрытии ридами. Эти данные были использованы для определения отношения уровня экспрессии EEJ, образующихся при удалении RI из изучаемых РНК и последующем воссоединении фланкирующих их экзонов, к уровню экспрессии одного из конститутивных EEJ соответствующего транскрипта. Предполагалось, что использование конститутивных EEJ, т. е. присутствующих в большинстве изоформ РНК отдельного гена, позволит оценить представленность полностью сплайсированных молекул в общем пуле транскриптов, принадлежащих целевым генам.

Результаты анализа, показанные на рис. 3, оказались сопоставимы с таковыми, полученными для клеток Kasumi-1. На этом рисунке EEJ, экспрессия которых сравнивалась между собой, обозначены буквой J и пронумерованы: первым указано альтернативное, а вторым – конститутивное событие (обоснование выбора конкретных EEJ для сравнения приведено в следующем разделе). Заслуживает внимания закономерность, прослеживающаяся как в случае клеток модельной линии, так и в случае первичных клеток ОМЛ и нормальных $CD34^+$ клеток: полное сплайсирование транскриптов исследуемых генов представляет собой редкое событие. По-видимому, основной вклад в пул РНК продуктов изучаемых генов вносят молекулы с RI. Причем это характерно как для лейкозных, так и для нормальных клеток крови человека. Однако можно отметить заметное выделение на общем фоне данных, полученных в отношении транскриптов гена *OGT* в нормальных $CD34^+$ клетках из красного костного мозга. В данном случае наблюдается относительно более высокая представленность РНК, не несущих RI.

Общий эффект кодирующих RI на структуру молекул мРНК и соответствующих белков. Кодировкой только новых полипептидов, очевидно, не исчерпывается влияние кодирующих RI на структуру молекул мРНК и соответствующих белков. Интерес представляет также вклад в общее разнообразие белков лейкозных клеток тех кодирующих RI, полипептиды которых уже известны и присутствуют в белковых базах данных. Для выяснения этого вопроса была проведена

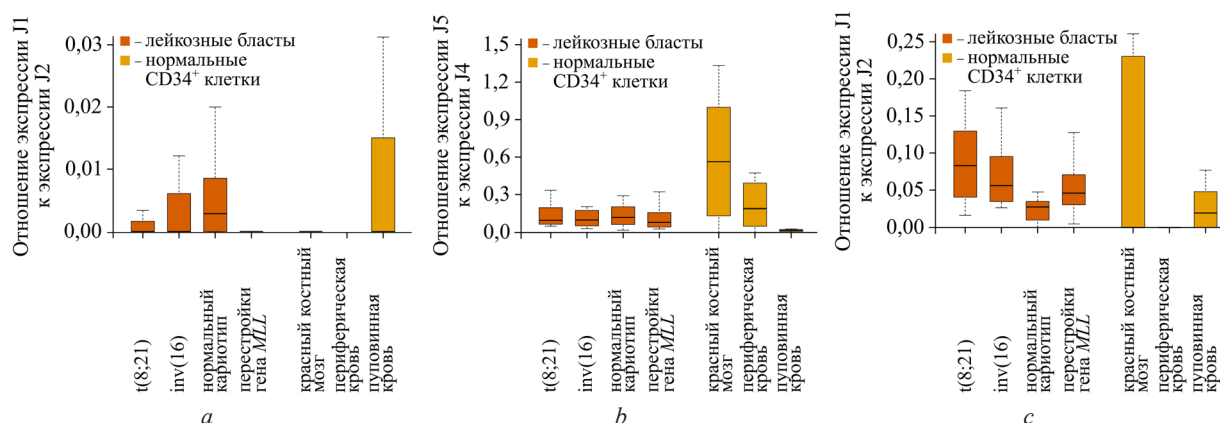


Рис. 3. Уровень экспрессии EEJ, свидетельствующих об удалении RI из транскриптов генов *UBE2M* (a), *OGT* (b) и *SLC10A3* (c), по отношению к уровню экспрессии соответствующих конститутивных EEJ в транскриптоме различных типов лейкозных и нормальных $CD34^+$ клеток крови человека

Fig. 3. Expression ratio between EEJ that supports RI splicing out and constitutive EEJ of *UBE2M* (a), *OGT* (b), *SLC10A3* (c) genes in the transcriptome of different types of human primary leukemia blasts or normal $CD34^+$ cells

реконструкция транскриптома клеток Kasumi-1 с помощью сборщика StringTie. После сборки осуществлялся поиск mORF в полноразмерных РНК и их трансляция *in silico* с помощью функционала R/Bioconductor пакета ORFhunter [28]. Далее определялось, на какую именно часть белка и каким именно образом оказывает влияние тот или иной кодирующий сохраненный интрон.

Анализ структуры молекул РНК, несущих кодирующие RI, показал, что сохраненные интроны могут быть разделены на 4 класса в зависимости от их локализации в пределах транскрипта и, как следствие, предполагаемого эффекта, оказываемого на структуру соответствующих белков: 1) RI, целиком локализованные в пределах 5'-нетранслируемой области транскрипта (от англ. untranslated region, UTR); 2) RI, затрагивающие как 5'-UTR, так и 5'-конец mORF; 3) RI, расположенные во внутренней части mORF; 4) RI, охватывающие и 3'-конец mORF, и 3'-UTR.

Примечательно, что интронами, расположенными в 5'-UTR (всего 3 RI), стали именно те, которые в предыдущем разделе были описаны нами как RI, кодирующие новые полипептидные последовательности. С целью определить, каким образом такие RI, находясь вне границ mORF, могут быть транслированы, нами был выполнен поиск в целевых транскриптах всех возможных ORF с использованием инструмента ORFfinder [29]. Идея базировалась на литературных данных, согласно которым в пределах одного транскрипта возможно существование нескольких ORF: главной и так называемых малых ORF (от англ. small open reading frames, smORF). Причем результаты последних исследований показывают, что триплет ATG выступает в качестве стартового кодона лишь у половины известных smORF [30]. По этой причине при решении описываемой задачи в качестве стартовых наряду с ATG рассматривались также альтернативные кодоны.

У гена *SLC10A3* сохранение интрона, расположенного между 1-м и 2-м экзонами, приводит к модификации структуры его 5'-UTR (рис. 4, а) и появлению smORF, кодирующей новый микробелок размером 109 а. о. (рис. 5, а), экспрессия которого подтверждается МС-данными. В случае гена *OGT* сохраняемый интрон располагается в кодирующей области и определяет инициацию процесса синтеза белка OGT с альтернативного нижележащего стартового кодона (рис. 4, б). Это, в свою очередь, обуславливает образование новой усеченной по N-концу изоформы данного белка. Кроме того, в пределах непосредственно интронной последовательности локализуется smORF, кодирующая новый микробелок размером 58 а. о. (рис. 5, б), экспрессия которого также подтверждается нашими МС-данными. Оба микробелка, согласно результатам моделирования, проведенного с помощью Conserved Domain Search и InterProScan [31, 32], содержат участки неупорядоченной структуры неизвестной функциональности. Наконец, в случае гена *UBE2M* сохранение интрона в мРНК приводит к N-удлинению полноразмерного белка на 115 а. о., причем данная белковая изоформа возникает благодаря инициации трансляции с альтернативного стартового кодона CTG внутри экзонной последовательности, предшествующей сохраняемому интрону (рис. 4, в). Опять же, эта новая изоформа белка UBE2M в пределах 115-аминокислотной N-концевой добавки содержит участок неупорядоченной структуры неизвестной функциональности (рис. 5, в).

Последовательности RI 2-го класса (всего 2 RI) содержат стартовый кодон трансляции и определяют N-конец белков. Гены, которым принадлежат такие RI, – *LMNB1* и *MOGS*. Первый кодирует ламин B1 – один из ламинов B-типа, вовлеченных в регуляцию организации хроматина, репликации и репарации ДНК, клеточного цикла, пролиферации и старения клеток [33]. Конечным продуктом экспрессии гена *MOGS* является фермент маннозил-олигосахарид глюкозидаза, необходимый для нормального протекания N-гликозилирования белков [34].

RI, вошедшие в класс интронов, локализованных в пределах mORF (всего 38 RI), обеспечивают синтез соответствующих внутренних фрагментов белков. Подобные кодирующие RI были обнаружены в РНК-продуктах, например, таких генов, как *BCLAF1* и *HNRNPL*. Исходно рассматриваемый только в роли регулятора транскрипции и клеточного апоптоза, BCLAF1 на сегодняшний день известен как белок, принимающий участие в широком спектре клеточных процессов, в том числе в процессинге и сплайсинге молекул пре-мРНК [35]. Продукт экспрессии *HNRNPL* – один из представителей белков класса гетерогенных ядерных рибонуклеопротеинов, способный влиять на протекание ряда биологических процессов, в частности на репарацию ДНК, альтернативный сплайсинг РНК, трансляцию и сигнальную трансдукцию [36].

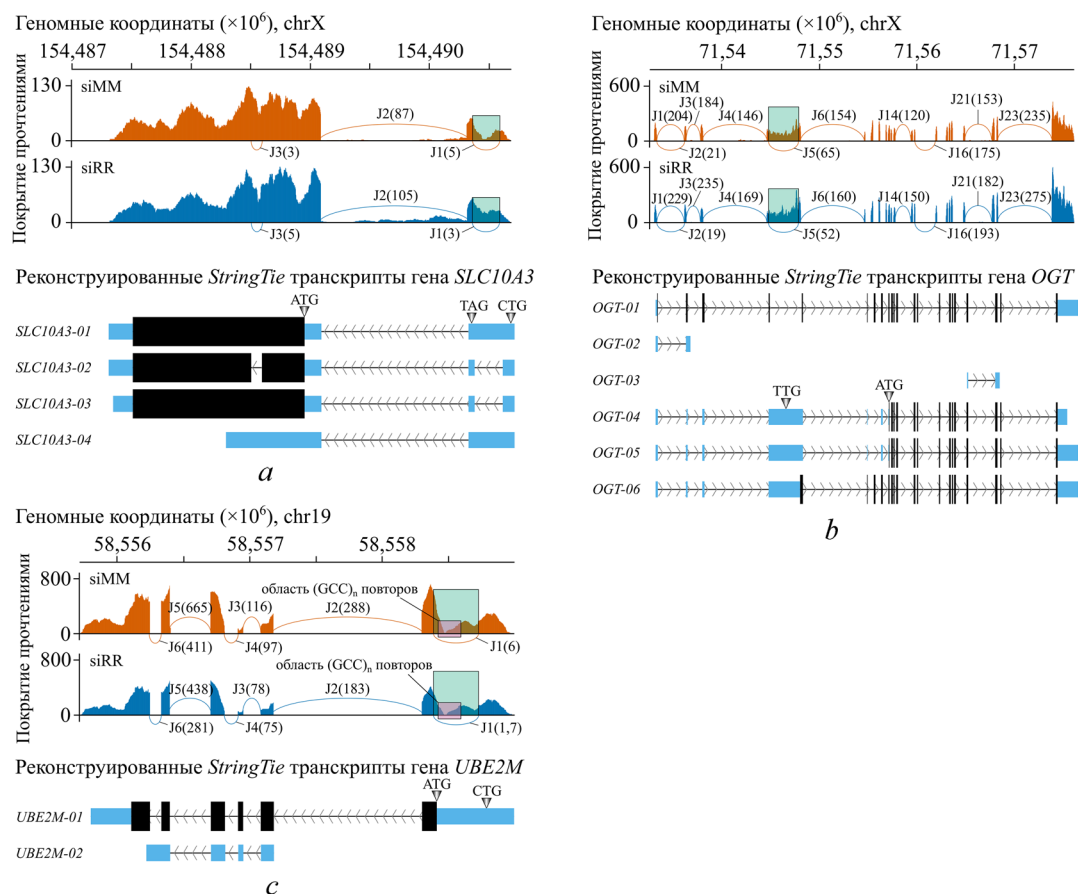


Рис. 4. Графы сплайсинга и структура реконструированных *StringTie* транскриптов генов *SLC10A3* (a), *OGT* (b) и *UBE2M* (c). Графы сплайсинга построены на основе данных секвенирования транскриптома клеток линии Kasumi-1 с нокдауном гибридного онкогена *RUNX1::RUNX1T1* и без него. Экзон-экзонные стыки отмечены латинской буквой J и пронумерованы. В скобках указано количество прочтений, подтверждающих конкретное сплайсинговое событие. Области сохранения интронов ограничены рамками

Fig. 4. Splicing graphs and *StringTie*-assembled transcripts of *SLC10A3* (a), *OGT* (b) and *UBE2M* (c) genes. Splicing graphs are based on total RNA-Seq data from *RUNX1::RUNX1T1* knockdown and control Kasumi-1 cells. Exon-exon junctions are labeled with the letter J and numbered. The number of reads supporting each splicing event is indicated in parentheses. Regions of retained introns are boxed

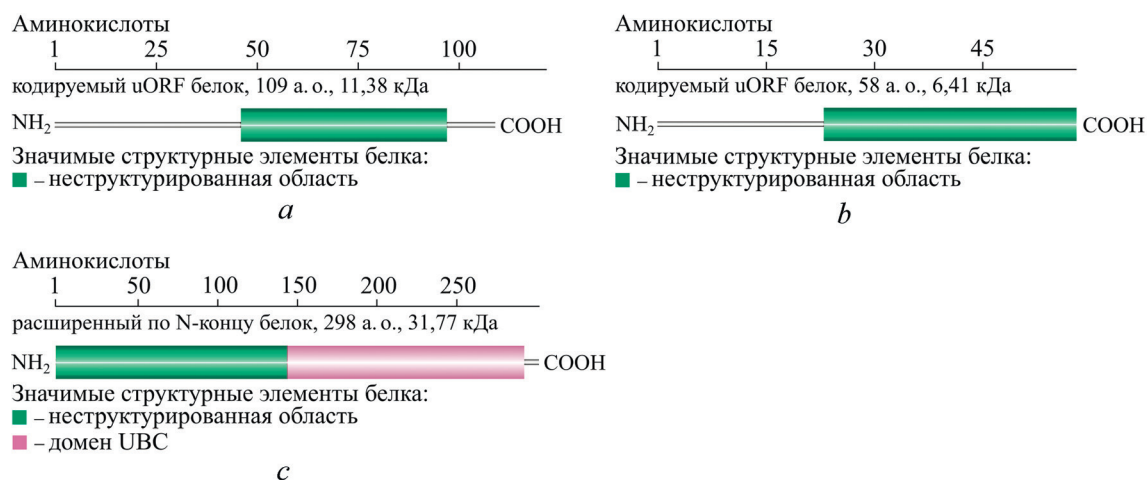


Рис. 5. Структура *in silico* транслированных ранее неизвестных белков, кодируемых генами *SLC10A3* (a), *OGT* (b) и *UBE2M* (c)

Fig. 5. *In silico* predicted structure of previously unknown proteins encoded by *SLC10A3* (a), *OGT* (b) and *UBE2M* (c) genes

RI, отнесенные к 4-му классу (всего 13 RI) несут стоп-кодон трансляции и детерминируют аминокислотную последовательность С-концевых участков белков. Среди RI указанного класса обнаружены, например, принадлежащие генам *NKRF* и *SH3BP1*. Ген *NKRF* кодирует ингибитор транскрипционного фактора NF- κ B, регулирующего экспрессию генов, чьи белковые продукты вовлечены в процессы воспаления, иммунного ответа, клеточной пролиферации и апоптоза [37]. Ген *SH3BP1* кодирует один из белков-регуляторов сигнального пути Ras [38]. Как видно, во всех этих случаях речь идет о генах, критически важных для жизни клеток.

Общий эффект некодирующих RI на структуру молекул мРНК и соответствующих белков. Количество RI, у которых в данном исследовании не было выявлено кодирующего потенциала, составило 3484 (или 98,4 % от общего списка идентифицированных RI). Из них 2239 могут быть локализованы в областях ORF белковых генов, описанных в базе данных Ensembl. Следовательно, такие RI могут влиять не только на структуру мРНК, но и на аминокислотную последовательность соответствующих белков.

Анализ полноразмерных РНК, содержащих некодирующие RI, выявил три потенциальных механизма влияния сохранения интронов на кодирующий потенциал таких молекул. Во-первых, при включении интрона в область ORF молекула мРНК может сохранять способность быть транслированной, но с использованием альтернативного нижележащего стартового кодона. Более подробно в качестве примера может быть рассмотрен ген *RASSF4* и его РНК/белковые продукты (рис. 6, а). Вследствие сохранения в транскриптах этого гена всей интронной последовательности, локализованной между экзонами 2 и 3, может образовываться усеченный по N-концу белок как результат использования нижележащего альтернативного стартового кодона. Косвенным подтверждением правомочности такого предположения служат протеомные МС-данные для клеток Kasumi-1, которые содержат МС-пептиды только части белка RASSF4, кодируемой экзонами гена, расположенными за сохраняемым интроном. Известно, что полноразмерный белковый продукт гена *RASSF4* является участником ряда сигнальных путей и функционирует в роли опухолевого

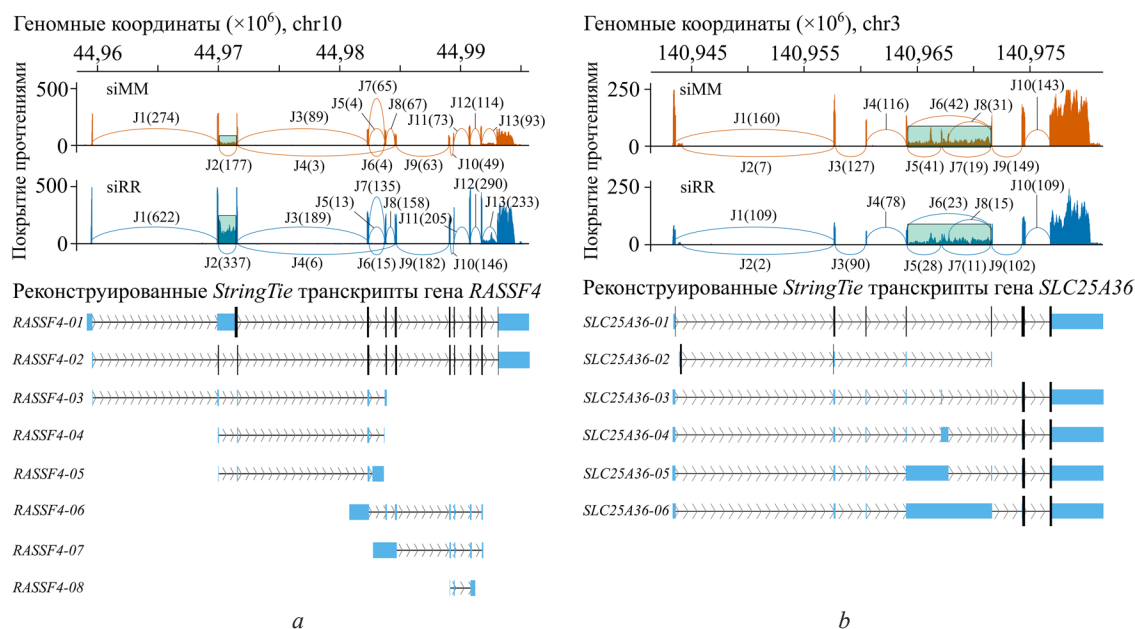


Рис. 6. Графы сплайсинга и структура реконструированных с помощью StringTie транскриптов генов *RASSF4* (а) и *SLC25A36* (б). Графы сплайсинга построены на основе данных секвенирования транскриптома клеток линии Kasumi-1 с нокдауном гибридного онкогена *RUNX1::RUNX1T1* и без него. Экзон-экзонные стыки отмечены латинской буквой J и пронумерованы. В скобках указано количество прочтений, подтверждающих конкретное сплайсинговое событие. Область сохранения интрона ограничена рамкой

Fig. 6. Splicing graphs and StringTie assembled transcripts of *RASSF4* (a) and *SLC25A36* (b) genes. Splicing graphs are based on total RNA-Seq data from *RUNX1::RUNX1T1* knockdown and control Kasumi-1 cells. Exon-exon junctions are designated by the letter J and numbered. The number of reads supporting each splicing event is indicated in parentheses. Region of retained intron is boxed

супрессора [39]. А вот какую роль может играть его N-усеченная изоформа, предстоит выяснить в будущем.

Второй механизм – фактически полная потеря кодирующего потенциала РНК-продуктов белкового гена из-за включения интрона. Примером является ген *SLC25A36* и его РНК/белковые продукты (рис. 6, *b*). Он кодирует митохондриальный транспортер пиримидиновых нуклеотидов, локализующийся во внутренней мембране митохондрий и обеспечивающий поддержание стабильности ДНК этих органелл и регуляцию мембранного потенциала [40]. В клетках Kasumi-1 мажорная фракция РНК-продуктов данного гена представлена молекулами, в ходе сплайсирования которых происходит сохранение интрона (целиком или фрагментарно), фланкированного экзонами 4 и 5. Воспользовавшись ассемблером StringTie мы провели реконструкцию полноразмерных транскриптов гена *SLC25A36* и идентифицировали в них mORF. Наше биоинформатическое моделирование показывает, что транскрипты, содержащие вышеупомянутый интрон, если и могут транслироваться, то только в короткие микробелки с нижележащего кодона ATG. Однако МС-данные не подтверждают экспрессии в клетках Kasumi-1 каких-либо белков гена *SLC25A36*, несмотря на высокую продукцию этим геном молекул РНК. Это, а также тот факт, что подобного рода микробелки характеризовались бы малой протяженностью и, по-видимому, не были бы способны нормально функционировать, свидетельствует в пользу обусловленной сохранением интрона утраты транскриптами гена *SLC25A36* кодирующего потенциала.

Наконец, возможен и третий механизм – появление РТС в молекулах РНК, несущих RI. Так, сборщик StringTie реконструировал 233 белок-кодирующих транскрипта, содержащих РТС и несущих в областях mORF RI (всего 175 RI), классифицированные как некодирующие. Очевидно, что транскрипты с РТС могут быть потенциальными мишенями для компонентов системы NMD. И даже если они транслируются, то из-за собственного невысокого присутствия в клетках (вследствие действия системы NMD) они не могут произвести такое количество белка, которого было бы достаточно для детекции с помощью использованной протеомной технологии.

Закключение. Таким образом, разработанный нами комбинированный подход, где использовались транскриптомные и протеомные данные, а также многошаговый биоинформатический конвейер для анализа таких данных, позволил провести углубленное изучение эффекта сохранения интронов на кодирующий потенциал РНК-продуктов белковых генов в модельных клетках ОМЛ человека. Полученные нами результаты позволяют разделить весь пул RI белковых генов на кодирующие и некодирующие. При этом на кодирующий потенциал РНК могут влиять представители обоих классов: кодирующие RI действуют напрямую, а некодирующие RI влияют опосредованно, через изменение структуры молекул РНК.

Вместе с тем результаты нашего исследования порождают и ряд новых вопросов:

каким образом обусловленное сохранением интрона усечение белка либо же приобретение им нового аминокислотного фрагмента отражается на его пространственной укладке, свойствах и функциональной активности вплоть до изменения фенотипа клетки;

обладают ли обнаруженные нами новые белки иммуногенностью и могут ли они рассматриваться как мишени для терапии лейкозов.

Эти и некоторые другие вопросы могут стать предметом дальнейших исследований.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Благодарности. Авторы признательны доктору Хосе Марин-Рубио из Института биологических наук при Ньюкаслском университете (г. Ньюкасл, Великобритания), а также профессору Олафу Хайденрайху из Центра детской онкологии имени Принцессы Максими (г. Утрехт, Нидерланды) за помощь в получении оригинальных протеомных данных для клеток Kasumi-1. Эта работа стала частью проектов «Разработка комплексного подхода по идентификации новых, зависящих от дифференциального сплайсинга, иммуногенных пептидов лейкозных клеток», выполняемого при финансовой поддержке Белорусского республиканского фонда фундаментальных

Acknowledgements. The authors thank Dr. Jose Luis Marin-Rubio (Biosciences Institute, Newcastle University, Newcastle-upon-Tyne, United Kingdom) as well as Professor Olaf Heidenreich (Princess Maxima Center for Pediatric Oncology, Utrecht, The Netherlands) for assistance in obtaining the original proteomic data for Kasumi-1 cells. The research was financially supported by the Belarusian Republican Foundation for Fundamental Research as part of the project “Development of a comprehensive approach to identify new differential splicing-dependent immunogenic peptides of leukemia cells” (grant No. M23-056, reg. No. 20231102). The described results were also obtained

исследований (грант № M23-056, № гос. регистрации 20231102), и «Разработка моделей, алгоритмов и программных средств для идентификации и аннотации сайтов генетического полиморфизма по данным геномного секвенирования», выполняемого в рамках подпрограммы «Междисциплинарные исследования и новые зарождающиеся технологии» ГПНИ «Конвергенция–2025» на 2021–2025 гг. (проект № 3.04.3.1, № гос. регистрации 20211918).

within the framework of the project “Development of models, algorithms and software for identification and annotation of genetic polymorphism sites based on genomic sequencing data” (State Program of Scientific Research “Convergence–2025”, Subprogram “Interdisciplinary research and new emerging technologies” (project No. 3.04.3.1, reg. No. 20211918).

Список использованных источников

1. The status of the human gene catalogue / P. Amaral, S. Carbonell-Sala, F. M. De La Vega [et al.] // *Nature*. – 2023. – Vol. 622, N 7981. – P. 41–47. <https://doi.org/10.1038/s41586-023-06490-x>
2. Ramanouskaya, T. V. The determinants of alternative RNA splicing in human cells / T. V. Ramanouskaya, V. V. Grinev // *Molecular Genetics and Genomics*. – 2017. – Vol. 292, N 6. – P. 1175–1195. <https://doi.org/10.1007/s00438-017-1350-0>
3. Liu, Q. Alternative splicing and isoforms: from mechanisms to diseases / Q. Liu, L. Fang, C. Wu // *Genes*. – 2022. – Vol. 13, N 3. – Art. 401. <https://doi.org/10.3390/genes13030401>
4. Intron retention as a mode for RNA-Seq data analysis / J.-T. Zheng, C.-X. Lin, Z.-Y. Fang, H.-D. Li // *Frontiers in Genetics*. – 2020. – Vol. 11. – Art. 586. <https://doi.org/10.3389/fgene.2020.00586>
5. Decoding of exon splicing patterns in the human RUNX1–RUNX1T1 fusion gene / V. V. Grinev, A. A. Migas, A. D. Kirsanova [et al.] // *The International Journal of Biochemistry and Cell Biology*. – 2015. – Vol. 68. – P. 48–58. <https://doi.org/10.1016/j.biocel.2015.08.017>
6. Differential fates of introns in gene expression due to global alternative splicing / A. Kumari, S. Sedehizadeh, J. D. Brook [et al.] // *Human Genetics*. – 2021. – Vol. 141. – P. 31–47. <https://doi.org/10.1007/s00439-021-02409-6>
7. Holding on to junk bonds: intron retention in cancer and therapy / G. Monteuiis, U. Schmitz, V. Petrova [et al.] // *Cancer Research*. – 2021. – Vol. 81, N 4. – P. 779–789. <https://doi.org/10.1158/0008-5472.can-20-1943>
8. RUNX1/RUNX1T1 mediates alternative splicing and reorganises the transcriptional landscape in leukemia / V. V. Grinev, F. Barneh, I. M. Ilyushonak [et al.] // *Nature Communications*. – 2021. – Vol. 12. – Art. 520. <https://doi.org/10.1038/s41467-020-20848-z>
9. The European Nucleotide Archive in 2022 / J. Burgin, A. Ahamed, C. Cummins [et al.] // *Nucleic Acids Research*. – 2023. – Vol. 51, N D1. – P. D121–D125. <https://doi.org/10.1093/nar/gkac1051>
10. Гринев, В. В. Сохранение интронов в транскриптом лейкозных и нормальных клеток крови человека / В. В. Гринев // *Молекулярная и прикладная генетика*. – 2018. – Т. 25. – С. 44–55.
11. Biostrings: Efficient manipulation of biological strings. – URL: <https://bioconductor.org/packages/Biostrings> (date of access: 12.09.2024).
12. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads / M. Pertea, G. M. Pertea, C. M. Antonescu [et al.] // *Nature Biotechnology*. – 2015. – Vol. 33. – P. 290–295. <https://doi.org/10.1038/nbt.3122>
13. Liao, Y. The R package *Rsubread* is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads / Y. Liao, G. K. Smyth, W. Shi // *Nucleic Acids Research*. – 2019. – Vol. 47, N 8. – P. e47. <https://doi.org/10.1093/nar/gkz114>
14. GitHub. – URL: <https://github.com/VGrinev/TranscriptomicFeatures> (date of access: 12.09.2024).
15. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput / V. Demichev, Ch. B. Messner, S. I. Vernardis [et al.] // *Nature Methods*. – 2020. – Vol. 17. – P. 41–44. <https://doi.org/10.1038/s41592-019-0638-x>
16. UniProt: the universal protein knowledgebase in 2023 / The UniProt Consortium // *Nucleic Acids Research*. – 2023. – Vol. 51, N D1. – P. D523–D531. <https://doi.org/10.1093/nar/gkac1052>
17. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation / Sh. Pujar, N. A. O’Leary, C. M. Farrell [et al.] // *Nucleic Acids Research*. – 2018. – Vol. 46, iss. D1. – P. D221–D228. <https://doi.org/10.1093/nar/gkx1031>
18. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation / N. A. O’Leary, M. W. Wright, J. R. Brister [et al.] // *Nucleic Acids Research*. – 2016. – Vol. 44, N D1. – P. D733–D745. <https://doi.org/10.1093/nar/gkv1189>
19. The comprehensive R archive network. – URL: <https://cran.r-project.org/> (date of access: 12.09.2024).
20. Bioconductor: Open source software for Bioinformatics. – URL: <https://www.bioconductor.org/> (date of access: 12.09.2024).
21. The Need for Guidelines in Publication of Peptide and Protein Identification Data / S. Carr, R. Aebersold, M. Baldwin [et al.] // *Molecular & Cellular Proteomics*. – 2004. – Vol. 3, N 6. – P. 531–533. <https://doi.org/10.1074/mcp.t400006-mcp200>
22. A Practical Guide to Small Protein Discovery and Characterization Using Mass Spectrometry / Ch. H. Ahrens, J. T. Wade, M. M. Champion, J. D. Langer // *Journal of Bacteriology*. – 2022. – Vol. 204, N 1. – Art. 00353. <https://doi.org/10.1128/jb.00353-21>
23. The Gene Ontology knowledgebase in 2023 / The Gene Ontology Consortium, S. A. Aleksander, J. Balhoff [et al.] // *Genetics*. – 2023. – Vol. 224, N 11. – Art. 31. <https://doi.org/10.1093/genetics/iyad031>
24. BLAST. – URL: <https://blast.ncbi.nlm.nih.gov/Blast.cgi> (date of access: 12.09.2024).
25. Protein neddylation and its role in health and diseases / Sh. Zhang, Q. Yu, Zh. Li [et al.] // *Signal Transduction and Targeted Therapy*. – 2024. – Vol. 9. – Art. 85. <https://doi.org/10.1038/s41392-024-01800-9>
26. OGT as potential novel target: Structure, function and inhibitors / N. Zhang, H. Jiang, K. Zhang [et al.] // *Chemico-Biological Interactions*. – Vol. 357. – Art. 109886. <https://doi.org/10.1016/j.cbi.2022.109886>
27. Claro da Silva T. The solute carrier family 10 (SLC10): Beyond bile acid transport / T. Claro da Silva, J. E. Polli, P. W. Swaan // *Molecular Aspects of Medicine*. – 2013. – Vol. 34, N 2–3. – P. 252–269. <https://doi.org/10.1016/j.mam.2012.07.004>

28. ORFhunter: An accurate approach to the automatic identification and annotation of open reading frames in human mRNA molecules / V. V. Grinev, M. M. Yatskou, V. V. Skakun [et al.] // *Software Impacts*. – 2022. – Vol. 12. – Art. 100268. <https://doi.org/10.1016/j.simpa.2022.100268>
29. Open Reading Frame Finder. – URL: <https://www.ncbi.nlm.nih.gov/orffinder/> (date of access: 12.09.2024).
30. A high-resolution map of human RNA translation / S. P. Chothani, E. Adami, A. A. Widjaja [et al.] // *Molecular Cell*. – 2022. – Vol. 82, N 15. – P. 2885–2899. <https://doi.org/10.1016/j.molcel.2022.06.023>
31. The conserved domain database in 2023 / J. Wang, F. Chitsaz, M. K. Derbyshire [et al.] // *Nucleic Acids Research*. – 2023. – Vol. 51, N D1. – P. D384–D388. <https://doi.org/10.1093/nar/gkac1096>
32. InterPro in 2022 / T. Paysan-Lafosse, M. Blum, S. Chuguransky [et al.] // *Nucleic Acids Research*. – 2023. – Vol. 51, N D1. – P. D418–D427. <https://doi.org/10.1093/nar/gkac993>
33. The wide and growing range of lamin B-related diseases: from laminopathies to cancer / C. Evangelisti, I. Rusciano, S. Mongiorgi [et al.] // *Cellular and Molecular Life Sciences*. – 2022. – Vol. 79. – Art. 126. <https://doi.org/10.1007/s00018-021-04084-2>
34. Congenital disorders of glycosylation type IIb with MOGS mutations cause early infantile epileptic encephalopathy, dysmorphic features, and hepatic dysfunction / R. Anzai, M. Tsuji, S. Yamashita [et al.] // *Brain and Development*. – 2021. – Vol. 43, N 3. – P. 402–410. <https://doi.org/10.1016/j.braindev.2020.10.013>
35. Function of BCLAF1 in human disease (Review) / Z. Yu, J. Zhu, H. Wang [et al.] // *Oncology Letters*. – 2022. – Vol. 23, N 2. – Art. 58. <https://doi.org/10.3892/ol.2021.13176>
36. Heterogeneous nuclear ribonucleoprotein (hnRNPL) in cancer / J. Gu, Zh. Chen, X. Chen, Zh. Wang // *Clinica Chimica Acta*. – 2020. – Vol. 507. – P. 286–294. <https://doi.org/10.1016/j.cca.2020.04.040>
37. Human NF- κ B repressing factor acts as a stress-regulated switch for ribosomal RNA processing and nucleolar homeostasis surveillance / M. Coccia, A. Rossi, A. Riccio [et al.] // *Proceedings of the National Academy of Sciences*. – 2017. – Vol. 114, N 5. – P. 1045–1050. <https://doi.org/10.1073/pnas.1616112114>
38. SH3BP1, an Exocyst-Associated RhoGAP, Inactivates Rac1 at the Front to Drive Cell Motility / M. C. Parrini, A. Sadou-Dubourgnoux, K. Aoki [et al.] // *Molecular Cell*. – 2011. – Vol. 42, N 5. – P. 650–661. <https://doi.org/10.1016/j.molcel.2011.03.032>
39. The role and function of Ras-association domain family in cancer: A review / M. R. Zinatizadeh, S. A. Momeni, P. K. Zarandi [et al.] // *Genes and Diseases*. – 2019. – Vol. 6, N 4. – P. 378–384. <https://doi.org/10.1016/j.gendis.2019.07.008>
40. The human SLC25A33 and SLC25A36 genes of solute carrier family 25 encode two mitochondrial pyrimidine nucleotide transporters / M. A. Di Noia, S. Todisco, A. Cirigliano [et al.] // *Journal of Biological Chemistry*. – 2014. – Vol. 289, N 48. – P. 33137–33148. <https://doi.org/10.1074/jbc.m114.610808>

References

1. Amaral P., Carbonell-Sala S., De La Vega F. M., Faial T., Frankish A., Gingeras T. [et al.]. The status of the human gene catalogue. *Nature*, 2023, vol. 622, no. 7981, pp. 41–47. <https://doi.org/10.1038/s41586-023-06490-x>
2. Ramanouskaya T. V., Grinev V. V. The determinants of alternative RNA splicing in human cells. *Molecular Genetics and Genomics*, 2017, vol. 292, no. 6, pp. 1175–1195. <https://doi.org/10.1007/s00438-017-1350-0>
3. Liu Q., Fang L., Wu C. Alternative splicing and isoforms: from mechanisms to diseases. *Genes*, 2022, vol. 13, no. 3, art. 401. <https://doi.org/10.3390/genes13030401>
4. Zheng J.-T., Lin C.-X., Fang Z.-Y., Li H.-D. Intron retention as a mode for RNA-Seq data analysis. *Frontiers in Genetics*, 2020, vol. 11, art. 586. <https://doi.org/10.3389/fgene.2020.00586>
5. Grinev V. V., Migas A. A., Kirsanova A. D., Mishkova O. A., Siomava N., Ramanouskaya T. V., Vaitsiankova A. V., Ilyushonak I. M., Nazarov P. V., Vallar L., Aleinikova O. V. Decoding of exon splicing patterns in the human RUNX1–RUNXIT1 fusion gene. *The International Journal of Biochemistry and Cell Biology*, 2015, vol. 68, pp. 48–58. <https://doi.org/10.1016/j.biocel.2015.08.017>
6. Kumari A., Sedehizadeh S., Brook J. D., Kozlowski P., Wojciechowska M. Differential fates of introns in gene expression due to global alternative splicing. *Human Genetics*, 2021, vol. 141, pp. 31–47. <https://doi.org/10.1007/s00439-021-02409-6>
7. Monteuuis G., Schmitz U., Petrova V., Kearney P. S., Rasko J. E. J. Holding on to junk bonds: intron retention in cancer and therapy. *Cancer Research*, 2021, vol. 81, no. 4, pp. 779–789. <https://doi.org/10.1158/0008-5472.can-20-1943>
8. Grinev V. V., Barnef F., Ilyushonak I. M., Nakjang S., Smink J., van Oort A. [et al.]. RUNX1/RUNXIT1 mediates alternative splicing and reorganizes the transcriptional landscape in leukemia. *Nature Communications*, 2021, vol. 12, art. 520. <https://doi.org/10.1038/s41467-020-20848-z>
9. Burgin J., Ahamed A., Cummins C., Devraj R., Gueye K., Gupta D., Gupta V. [et al.]. The European Nucleotide Archive in 2022. *Nucleic Acids Research*, 2023, vol. 51, no. D1, pp. D121–D125. <https://doi.org/10.1093/nar/gkac1051>
10. Grinev V. V. Intron retention in the transcriptome of leukemic and normal human blood cells. *Molekulyarnaya i prikladnaya genetika* [Molecular and Applied Genetics], 2018, vol. 25, pp. 45–56 (in Russian).
11. *Biostrings: Efficient manipulation of biological strings*. Available at: <https://bioconductor.org/packages/Biostrings> (accessed 12.09.2024).
12. Pertea M., Pertea G. M., Antonescu C. M., Chang T.-Ch., Mendell J. T., Salzberg S. L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 2015, vol. 33, pp. 290–295. <https://doi.org/10.1038/nbt.3122>
13. Liao Y., Smyth G. K., Shi W. The R package *Rsubread* is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, 2019, vol. 47, no. 8, p. e47. <https://doi.org/10.1093/nar/gkz114>
14. *GitHub*. Available at: <https://github.com/VGrinev/TranscriptomicFeatures/> (accessed 12.09.2024).
15. Demichev V., Messner Ch. B., Vernardis S. I., Lilley K. S., Ralser M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 2020, vol. 17, pp. 41–44. <https://doi.org/10.1038/s41592-019-0638-x>
16. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 2023, vol. 51, no. D1, pp. D523–D531. <https://doi.org/10.1093/nar/gkac1052>

17. Pujar Sh., O'Leary N. A., Farrell C. M., Loveland J. E., Mudge J. M., Wallin C., Girón C. G. [et al.]. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Research*, 2018, vol. 46, no. D1, pp. D221–D228. <https://doi.org/10.1093/nar/gkx1031>
18. O'Leary N. A., Wright M. W., Brister J. R., Ciufio S., Haddad D., McVeigh R., Rajput B. [et al.]. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 2016, vol. 44, no. D1, pp. D733–D745. <https://doi.org/10.1093/nar/gkv1189>
19. *The comprehensive R archive network*. Available at: <https://cran.r-project.org/> (accessed 12.09.2024).
20. *Bioconductor: Open source software for Biobinformatics*. Available at: <https://www.bioconductor.org/> (accessed 12.09.2024).
21. Carr S., Aebersold R., Baldwin M., Burlingame A., Clauser K., Nesvizhskii A. The Need for Guidelines in Publication of Peptide and Protein Identification Data. *Molecular & Cellular Proteomics*, 2004, vol. 3, no. 6, pp. 531–533. <https://doi.org/10.1074/mcp.t400006-mcp200>
22. Ahrens Ch. H., Wade J. T., Champion M. M., Langer J. D. A Practical Guide to Small Protein Discovery and Characterization Using Mass Spectrometry. *Journal of Bacteriology*, 2022, vol. 204, no. 1, art. 00353. <https://doi.org/10.1128/jb.00353-21>
23. The Gene Ontology Consortium, Aleksander S. A., Balhoff J., Carbon S., Cherry J. M., Drabkin H. J., Ebert D. The Gene Ontology knowledgebase in 2023. *Genetics*, 2023, vol. 224, no. 11, art 31. <https://doi.org/10.1093/genetics/iyad031>
24. *BLAST*. Available at: <https://blast.ncbi.nlm.nih.gov/Blast.cgi> (accessed 12.09.2024).
25. Zhang Sh., Yu Q., Li Zh., Zhao Y., Sun Y. Protein neddylation and its role in health and diseases. *Signal Transduction and Targeted Therapy*, 2024, vol. 9, art. 85. <https://doi.org/10.1038/s41392-024-01800-9>
26. Zhang N., Jiang H., Zhang K., Zhu J., Wang Zh., Long Y., He Y., Feng F., Liu W., Ye F., Qu W. OGT as potential novel target: Structure, function and inhibitors. *Chemico-Biological Interactions*, vol. 357, art. 109886. <https://doi.org/10.1016/j.cbi.2022.109886>
27. Claro da Silva T., Polli J. E., Swaan P. W. The solute carrier family 10 (SLC10): Beyond bile acid transport. *Molecular Aspects of Medicine*, 2013, vol. 34, no. 2–3, pp. 252–269. <https://doi.org/10.1016/j.mam.2012.07.004>
28. Grinev V. V., Yatskou M. M., Skakun V. V., Chepeleva M. K., Nazarov P. V. ORFhunter: An accurate approach to the automatic identification and annotation of open reading frames in human mRNA molecules. *Software Impacts*, 2022, vol. 12, art. 100268. <https://doi.org/10.1016/j.simpa.2022.100268>
29. *Open Reading Frame Finder*. Available at: <https://www.ncbi.nlm.nih.gov/orffinder/> (accessed 12.09.2024).
30. Chothani S. P., Adami E., Widjaja A. A., Langley S. R., Viswanathan S., Pua Ch. J. [et al.]. A high-resolution map of human RNA translation. *Molecular Cell*, 2022, vol. 82, no. 15, pp. 2885–2899. <https://doi.org/10.1016/j.molcel.2022.06.023>
31. Wang J., Chitsaz F., Derbyshire M. K., Gonzales N. R., Gwadz M., Lu Sh. [et al.]. The conserved domain database in 2023. *Nucleic Acids Research*, 2023, vol. 51, no. D1, pp. D384–D388. <https://doi.org/10.1093/nar/gkac1096>
32. Paysan-Lafosse T., Blum M., Chuguransky S., Grego T., Pinto B. L., Salazar G. A., Bileschi M. L. [et al.]. InterPro in 2022. *Nucleic Acids Research*, 2023, vol. 51, no. D1, pp. D418–D427. <https://doi.org/10.1093/nar/gkac993>
33. Evangelisti C., Rusciano I., Mongiorgi S., Ramazzotti G., Lattanzi G., Manzoli L., Cocco L., Ratti S. The wide and growing range of lamin B-related diseases: from laminopathies to cancer. *Cellular and Molecular Life Sciences*, 2022, vol. 79, art. 126. <https://doi.org/10.1007/s00018-021-04084-2>
34. Anzai R., Tsuji M., Yamashita S., Wada Y., Okamoto N., Saitsu H., Matsumoto N., Goto T. Congenital disorders of glycosylation type IIb with MOGS mutations cause early infantile epileptic encephalopathy, dysmorphic features, and hepatic dysfunction. *Brain and Development*, 2021, vol. 43, no. 3, pp. 402–410. <https://doi.org/10.1016/j.braindev.2020.10.013>
35. Yu Z., Zhu J., Wang H., Li H., Jin X. Function of BCLAF1 in human disease (Review). *Oncology Letters*, 2022, vol. 23, no. 2, art. 58. <https://doi.org/10.3892/ol.2021.13176>
36. Gu J., Chen Zh., Chen X., Wang Zh. Heterogeneous nuclear ribonucleoprotein (hnRNPL) in cancer. *Clinica Chimica Acta*, 2020, vol. 507, pp. 286–294. <https://doi.org/10.1016/j.cca.2020.04.040>
37. Coccia M., Rossi A., Riccio A., Trotta E., Santoro M. G. Human NF-κB repressing factor acts as a stress-regulated switch for ribosomal RNA processing and nucleolar homeostasis surveillance. *Proceedings of the National Academy of Sciences*, 2017, vol. 114, no. 5, pp. 1045–1050. <https://doi.org/10.1073/pnas.1616112114>
38. Parrini M. C., Sadou-Dubourgoux A., Aoki K., Kunida K., Biondini M., Hatzoglou A., Poulet P., Formstecher E., Yeaman Ch., Matsuda M., Rossé C., Camonis J. SH3BP1, an Exocyst-Associated RhoGAP, Inactivates Rac1 at the Front to Drive Cell Motility. *Molecular Cell*, 2011, vol. 42, no. 5, pp. 650–661. <https://doi.org/10.1016/j.molcel.2011.03.032>
39. Zinatizadeh M. R., Momeni S. A., Zarandi P. K., Chalbatani G. M., Dana H., Mirzaei H. R., Akbari M. E., Miri S. R. The Role and Function of Ras-association domain family in Cancer: A Review. *Genes and Diseases*, 2019, vol. 6, no. 4, pp. 378–384. <https://doi.org/10.1016/j.gendis.2019.07.008>
40. Di Noia M. A., Todisco S., Cirigliano A., Rinaldi T., Agrimi G., Iacobazzi V., Palmieri F. The Human SLC25A33 and SLC25A36 genes of solute carrier family 25 encode two mitochondrial pyrimidine nucleotide transporters. *Journal of Biological Chemistry*, vol. 289, no. 48, pp. 33137–33148. <https://doi.org/10.1074/jbc.m114.610808>

Інфармацыя аб аўтарах

Гузова Екатерина Витальевна – аспирант, мл. науч. сотрудник. Белорусский государственный университет (пр. Независимости, 4, 220030, г. Минск, Республика Беларусь). E-mail: k.guzowwa0099@gmail.com

Гавричков Александрович – студент. Белорусский государственный университет (пр. Независимости, 4, 220030, г. Минск, Республика Беларусь). E-mail: gavrichkow@gmail.com

Information about the authors

Katsiaryna V. Huzava – Postgraduate Student, Junior Researcher. Belarusian State University (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: k.guzowwa0099@gmail.com

Aleksey A. Gavrichkov – Student. Belarusian State University (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: gavrichkow@gmail.com

Ильющёнок Илья Николаевич – канд. биол. наук, доцент. Белорусский государственный университет (пр. Независимости, 4, 220030, г. Минск, Республика Беларусь). E-mail: nov.ilyushonok@gmail.com

Романовская Татьяна Владимировна – канд. биол. наук, доцент. Белорусский государственный университет (пр. Независимости, 4, 220030, г. Минск, Республика Беларусь). E-mail: romanovskaya.t@mail.ru

Лиховец Максим Александрович – магистрант. Белорусский государственный университет (пр. Независимости, 4, 220030, г. Минск, Республика Беларусь). E-mail: likhovetsmax@gmail.com

Скакун Виктор Васильевич – канд. физ.-мат. наук, доцент. Белорусский государственный университет (пр. Независимости, 4, 220030, г. Минск, Республика Беларусь). E-mail: Skakun@bsu.by

Яцков Николай Николаевич – канд. физ.-мат. наук, доцент, заведующий кафедрой. Белорусский государственный университет (пр. Независимости, 4, 220030, г. Минск, Республика Беларусь). E-mail: Yatskou@bsu.by

Гринец Василий Викторович – канд. биол. наук, доцент. Белорусский государственный университет (пр. Независимости, 4, 220030, г. Минск, Республика Беларусь). E-mail: grinev_vv@bsu.by

Ilya M. Ilyushonak – Ph. D. (Biol.), Associate Professor. Belarusian State University (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: nov.ilyushonok@gmail.com

Tatsiana V. Ramanouskaya – Ph. D. (Biol.), Associate Professor. Belarusian State University (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: romanovskaya.t@mail.ru

Maksim A. Likhavets – Master Student. Belarusian State University (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: likhovetsmax@gmail.com

Victor V. Skakun – Ph. D. (Math.), Associate Professor. Belarusian State University (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: Skakun@bsu.by

Mikalai M. Yatskou – Ph. D. (Math.), Associate Professor, Head of the Department. Belarusian State University (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: Yatskou@bsu.by

Vasily V. Grinev – Ph. D. (Biol.), Associate Professor. Belarusian State University (4, Nezavisimosti Ave., 220030, Minsk, Republic of Belarus). E-mail: grinev_vv@bsu.by