

---

# ГЕНЕТИКА И МОЛЕКУЛЯРНАЯ БИОЛОГИЯ

---

## GENETICS AND MOLECULAR BIOLOGY

---

УДК 577.214.5

### ВКЛАД РАЗЛИЧНЫХ МЕХАНИЗМОВ ГЕНЕРАЦИИ АЛЬТЕРНАТИВНЫХ ТРАНСКРИПТОВ В РАЗНООБРАЗИЕ мРНК ГИБРИДНОГО ОНКОГЕНА *RUNX1-RUNX1T1* ЧЕЛОВЕКА

**И. Н. ИЛЬЮШЁНОК<sup>1)</sup>, А. А. МИГАС<sup>2)</sup>, А. Ю. СУХАРЕВСКИЙ<sup>3)</sup>,  
О. Д. ШНАЙДЕР<sup>4)</sup>, В. В. ГРИНЕВ<sup>1)</sup>**

<sup>1)</sup>Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь

<sup>2)</sup>Республиканский научно-практический центр детской онкологии, гематологии и иммунологии,  
ул. Фрунзенская, 43, 223053, д. Боровляны, Боровлянский с/с, Минский район, Беларусь

<sup>3)</sup>Институт клеточной биологии, Эдинбургский университет, Роджер-ленд-билдингс,  
Александр-Крам-Браун-роуд, EH9 3FF, г. Эдинбург, Великобритания

<sup>4)</sup>Университетский госпиталь Карла Густава Каруса, Дрезденский технический университет,  
Моммзенитрассе, 9, 01069, г. Дрезден, Германия

---

#### Образец цитирования:

Ильюшёнков И.Н., Мигас А.А., Сухаревский А.Ю., Шнайдер О.Д., Гринев В.В. Вклад различных механизмов генерации альтернативных транскриптов в разнообразие мРНК гибридного онкогена *RUNX1-RUNX1T1* человека. *Журнал Белорусского государственного университета. Биология*. 2019;2:45–59.  
<https://doi.org/10.33581/2521-1722-2019-2-45-59>

#### For citation:

Ilyushonak IM, Migas AA, Sukhareuski AY, Schneider AD, Grinev VV. The contribution of various mechanisms to mRNA diversity of human fusion oncogene *RUNX1-RUNX1T1*. *Journal of the Belarusian State University. Biology*. 2019;2:45–59. Russian.  
<https://doi.org/10.33581/2521-1722-2019-2-45-59>

---

#### Авторы:

**Илья Николаевич Ильюшёнков** – ассистент кафедры генетики биологического факультета.

**Александр Александрович Мигас** – старший научный сотрудник лаборатории иммунологических исследований.

**Андрей Юрьевич Сухаревский** – аспирант. Научный руководитель – кандидат биологических наук С. В. Маковец.

**Оксана Дмитриевна Шнайдер** – кандидат биологических наук; научный сотрудник подразделения медицинской и системной биологии медицинского факультета.

**Василий Викторович Гринев** – кандидат биологических наук, доцент; доцент кафедры генетики биологического факультета.

#### Authors:

**Ilya M. Ilyushonak**, assistant at the department of genetics, faculty of biology.

[nov.ilyushonok@gmail.com](mailto:nov.ilyushonok@gmail.com)

<http://orcid.org/0000-0002-5377-366X>

**Aleksandr A. Migas**, senior researcher at the laboratory of immunological research.

[alexandr.migas@gmail.com](mailto:alexandr.migas@gmail.com)

**Andrei Y. Sukhareuski**, postgraduate student.

[andrey\\_suh@tut.by](mailto:andrey_suh@tut.by)

**Aksana D. Schneider**, PhD (biology); researcher at the section of medical systems biology, medical faculty.

[aksana.schneider@tu-dresden.de](mailto:aksana.schneider@tu-dresden.de)

**Vasily V. Grinev**, PhD (biology), docent; associate professor at the department of genetics, faculty of biology.

[grinev\\_vv@bsu.by](mailto:grinev_vv@bsu.by)

На большой выборке экзонов гибридного онкогена *RUNX1-RUNX1T1*, полученных из различных источников, исследованы закономерности генерации его альтернативных транскриптов. Показано, что альтернативные экзоны в абсолютном большинстве случаев образуются путем модификации канонических экзонов; транскрипты, включающие такие экзонные варианты, имеют очень низкий уровень экспрессии. Установлено, что около 80 % альтернативных экзонных вариантов генерируются «горячими областями», к которым относятся экзоны 4a, 6, 8b, 9, 11 и 12. Обнаружена новая точка инициации транскрипции, лежащая в области предсказанного биоинформатически промотора. Также установлено, что в лейкозных клетках коэкспрессируются транскрипты гибридного онкогена как с полноразмерными, так и с укороченными 3'-нетранслируемыми областями.

**Ключевые слова:** гибридный онкоген *RUNX1-RUNX1T1*; альтернативный сплайсинг РНК; высокопроизводительное секвенирование; биологический шум.

**Благодарность.** Работа выполнена в рамках подпрограммы «Объединение» государственной программы научных исследований «Конвергенция-2020» (задание 3.08.03 (№ 469/54)).

## THE CONTRIBUTION OF VARIOUS MECHANISMS TO mRNA DIVERSITY OF HUMAN FUSION ONCOGENE *RUNX1-RUNX1T1*

I. M. ILYUSHONAK<sup>a</sup>, A. A. MIGAS<sup>b</sup>, A. Y. SUKHAREUSKI<sup>c</sup>, A. D. SCHNEIDER<sup>d</sup>, V. V. GRINEV<sup>a</sup>

<sup>a</sup>Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus

<sup>b</sup>Belarusian Research Center for Pediatric Oncology, Hematology and Immunology,  
43 Frunzenskaja Street, Baraŭliany 223053, Baraŭlianski sielsaviet, Minsk region, Belarus

<sup>c</sup>Institute of Cell Biology, University of Edinburgh, Roger Land Building,  
Alexander Crum Brown Road, EH9 3FF, Edinburgh, United Kingdom

<sup>d</sup>University Hospital Carl Gustav Carus, Dresden University of Technology,  
9 Mommsenstrasse, Dresden 01069, Germany

Corresponding author: I. M. Ilyushonak (nov.ilyushonok@gmail.com)

In this work, we used a comprehensive set of the fusion oncogene *RUNX1-RUNX1T1* alternative exons to analyze the patterns of its mRNA generation. We found that the waste majority of alternative exons are modified variants of canonical exons, and the transcripts, including such exons, have a very low expression level. The «hot regions», including exons 4a, 6, 8b, 9, 11 and 12, produces about 80 % of such variants. Also we described a new transcription start region of *RUNX1-RUNX1T1* and provide the evidences of co-expression of the fusion RNAs with normal and shortened 3'-UTRs in leukemic cells.

**Keywords:** fusion oncogene *RUNX1-RUNX1T1*; alternative RNA splicing; high throughput sequencing; biological noise.

**Acknowledgements.** This work is a part of sub-programme «Union» of the State Program on Scientific Research «Convergence-2020» (grant 3.08.03 (No. 469/54)).

### Введение

Транслокация t(8; 21)(q22; q22) (далее по тексту – t(8; 21)), в процессе которой хромосомы 8 и 21 человека обмениваются участками, приводит к образованию двух деривативных хромосом – der8 и der21. Этот процесс сопряжен с объединением двух неродственных геномных локусов, кодирующих белки RUNX1 (хромосома 21) и RUNX1T1 (хромосома 8). Два гибридных гена, образующихся путем такого объединения, называются *RUNX1-RUNX1T1* и *RUNX1T1-RUNX1*. Последний локализован на der21 и является молчащим [1]. Белок RUNX1-RUNX1T1 есть транскрипционный фактор, вовлеченный в развитие острого миелоидного лейкоза (ОМЛ), однако его роль в этом процессе, несмотря на почти 25-летнюю историю изучения, не до конца ясна.

Гибридный онкоген *RUNX1-RUNX1T1* служит источником большого количества альтернативных транскриптов [2; 3], однако общие закономерности и глобальные правила, управляющие их образованием, неочевидны. Настоящее исследование является продолжением наших предыдущих работ [3; 4]. Его цель – максимально полное описание возможных альтернативных вариантов мРНК гибридного онкогена и оценка вклада в их разнообразие трех основных механизмов: 1) использования альтернативных промоторов; 2) альтернативного сплайсинга; 3) применения альтернативных точек полиаденилирования.

### Материалы и методы

**Клеточные линии и биоматериал пациентов.** Клетки модельной линии ОМЛ Kasumi-1 инкубировали при температуре +37 °C во влажной атмосфере с 5 % CO<sub>2</sub> в среде RPMI-1640 (Sigma-Aldrich Co.,

США) по стандартной методике. Мононуклеары костного мозга 12 пациентов с t(8;21)-положительной формой ОМЛ, проходящих лечение в Республиканском научно-практическом центре детской онкологии, гематологии и иммунологии, были получены при помощи Histopaque (*Sigma-Aldrich*, США). Манипуляции с материалом пациентов выполнялись в соответствии с рекомендациями Хельсинкской декларации и были одобрены этическими комиссиями организаций-участников.

**Выделение тотальной РНК, синтез комплементарной ДНК (кДНК), ПЦР с обратной транскрипцией и количественная ПЦР.** Выделение тотальной РНК осуществлялось с использованием TRI Reagent<sup>TM</sup> (*Sigma-Aldrich*, США) в соответствии с рекомендациями производителя; кДНК для секвенирования по Сэнгеру и протоколу прицельного RNA-Seq синтезировалась при помощи обратной транскриптазы М-MLV (*Thermo Fisher*, США); для верификации экзонных стыков и количественной ПЦР использовалась кДНК, синтезированная при помощи обратной транскриптазы RevAid (*Thermo Fisher*, США). В качестве затравок применялись рандомные гексамеры (группа данных *кДНК*, количественная ПЦР) либо sp-oligo(dT)-праймер (группа данных *прицельный RNA-Seq*). ПЦР с обратной транскрипцией (ОТ-ПЦР) проводилась по стандартной методике в амплификаторе BioRad C1000 при помощи ДНК-полимеразы DreamTaq (*Thermo Fisher*, США). Количественная ПЦР проводилась в амплификаторе BioRad CFX96 с использованием Maxima SYBR Green qPCR Master Mix (*Thermo Fisher*, США). Нормализация экспрессии осуществлялась по отношению к экспрессии гена *TBP* [4], расчет производился по формуле Пфаффа [5], для вычисления эффективности амплификации применялась программа *LinReg PCR* [6; 7].

**Создание библиотеки кДНК.** ПЦР-продукты, синтезированные на матрице кДНК лейкозных клеток, лигировались в вектор pTZ57R/T, который использовался для трансформации клеток *E. coli* XL-1 Blue. Рекombинантная ДНК выделялась из клеток при помощи GeneJET Plasmid Miniprep Kit (*Thermo Fisher*, США) и секвенировалась.

**Прицельный RNA-Seq.** Массовое секвенирование фрагментов кДНК выполняли по технологии парного прицельного (направленного) RNA-Seq. Фрагменты РНК в кДНК-форме готовились на основе ПЦР-ампликонов кДНК гибридного онкогена. Создание библиотеки кДНК осуществлялось при помощи Nextera DNA Library Preparation Kit (*Illumina*, США), секвенирование проводилось с использованием Illumina MiSeq (*Illumina*, США) в соответствии с рекомендациями фирмы-производителя. Для каждого образца генерировалось по два файла формата FASTQ: по одному файлу на прямое и обратное прочтение. Удаление последовательностей адаптеров осуществлялось при помощи программы *Trimmomatic* v.0.32 [8].

**Быстрая амплификация 3'-концов кДНК (англ. *rapid amplification of cDNA 3'-ends*, 3'-RACE).** При синтезе кДНК в качестве затравки использовался особый праймер sp-oligo(dT), содержащий поли-Т-участок с двумя вырожденными нуклеотидами на 3'-конце и олигонуклеотидную навеску на 5'-конце. В дальнейшем эта кДНК подвергалась амплификации с применением обратного праймера, соответствующего олигонуклеотидной навеске sp-oligo(dT)-праймера. Полученные ампликоны секвенировались по протоколу *прицельного RNA-Seq*, после чего адаптерные последовательности удалялись и данные анализировались при помощи программы *KLEAT-2.0* [9].

**Гель-электрофорез ДНК.** ПЦР-ампликоны разделялись в 1–2 % агарозном геле и очищались при помощи QIAquick Gel Extraction Kit (*Qiagen*, Германия).

**Картирование сиквенсов РНК.** Сиквенс точки рекомбинации между хромосомами 8 и 21, описанный для Kasumi-1 в [10; 11], был картирован на сборку генома человека GRCh38/hg38 с помощью программного инструмента *BLAT* [12]. Это позволило определить координаты областей хромосом 8 и 21, входящих в состав der8, после чего соответствующие фрагменты хромосом были объединены в единый сиквенс, сохраненный в формате FASTA. Нуклеотидные последовательности РНК картировались на der8 с применением установленной локально версии *BLAT*. Полученные файлы формата PSL конвертировались в формат GTF геномным браузером UCSC [13].

**Секвенирование по методу Сэнгера** осуществлялось при помощи *BigDye v.3.1 Terminator Cycle sequencing Kit* (*Applied Biosystems*, США) в соответствии с рекомендациями производителя.

**Реконструкция мРНК из библиотеки чтений RNA-Seq.** Для реконструкции полноразмерных мРНК использовались библиотеки чтений – коротких фрагментов РНК, нуклеотидная последовательность которых была определена с помощью высокопроизводительного секвенирования. Предобработанные чтения в формате FASTQ, загруженные из репозитория European Nucleotide Archive (исследование PRJNA236604), включали три независимых экспериментальных повтора для клеток, обработанных либо неактивными, либо *анти-RUNX1-RUNXIT1* специфичными короткими интерферирующими РНК. Методика реконструкции полноразмерных мРНК описана нами ранее в [14]. Данные *прицельного RNA-Seq* обрабатывались аналогичным образом, но без установки порогов минимального покрытия экзон-экзонных стыков чтениями.

**Биоинформатическая оценка уровня экспрессии.** Прямой подсчет чтений, покрывающих экзон-ные стыки, осуществлялся при помощи Sashimi-плотов, построенных в программе *IGV* [15]. Для оценки с помощью *Salmon* [16] сиквенсы полноразмерных транскриптов экстрагировались из GTF-файлов при помощи утилиты *gffread*, после чего передавались локально установленной версии *Salmon* вместе с исходными файлами FASTQ. Анализ трех экспериментальных повторов RNA-Seq осуществлялся одним проходом программы. Полученные значения, выраженные в транскриптах на миллион молекул мРНК (англ. *transcripts per million*, TPM), трансформировались в проценты от общей доли пула гибридной мРНК.

**Биоинформатическая идентификация артефактов обратной транскрипции.** Оценка идентичности потенциальных сайтов микрогомологии проводилась при помощи программных пакетов *Biostrings* [17], *Rsamtools* [18] и *GenomicFeatures* [19]. Анализ вторичных структур и свободной энергии Гиббса осуществлялся при помощи *ViennaRNA Package* [20].

**Биоинформатическая идентификация мод альтернативного сплайсинга.** Алгоритм классификации сплайсинговых событий был описан нами ранее [21].

### Общая характеристика набора данных

В рамках данного исследования использовались 3 набора данных. Первый набор объединяет РНК, секвенированные по методу Сэнгера. Он включает 135 полноразмерных транскриптов гибридного онкогена [3], 55 коротких экспрессированных последовательностей (англ. *expressed sequence tag*, EST) из образцов костного мозга пациентов с t(8; 21)-положительной формой ОМЛ и 14 коротких EST, обнаруженных в транскрипте клеточной линии Kasumi-1. В дальнейшем будем обозначать эту группу данных как **кДНК**. Второй набор включает транскрипты гибридного онкогена, собранные из данных полного прочтения транскриптома клеток модельной линии ОМЛ Kasumi-1, описанные в [14] (**полный RNA-Seq**), а третий – ПЦР-продукты (включая результаты 3'-RACE), полученные на кДНК клеток линии Kasumi-1 и прочитанные при помощи RNA-Seq (**прицельный RNA-Seq**).

Указанные РНК включают 316 уникальных экзон-экзонных стыков (англ. *exon-exon junction*, EEJs). Как можно видеть на рис. 1, перекрытие между EEJs в группах **кДНК** и **RNA-Seq** незначительно. При этом 70 % всех сайтов сплайсинга локализовано внутри канонических экзонов гибридного онкогена,

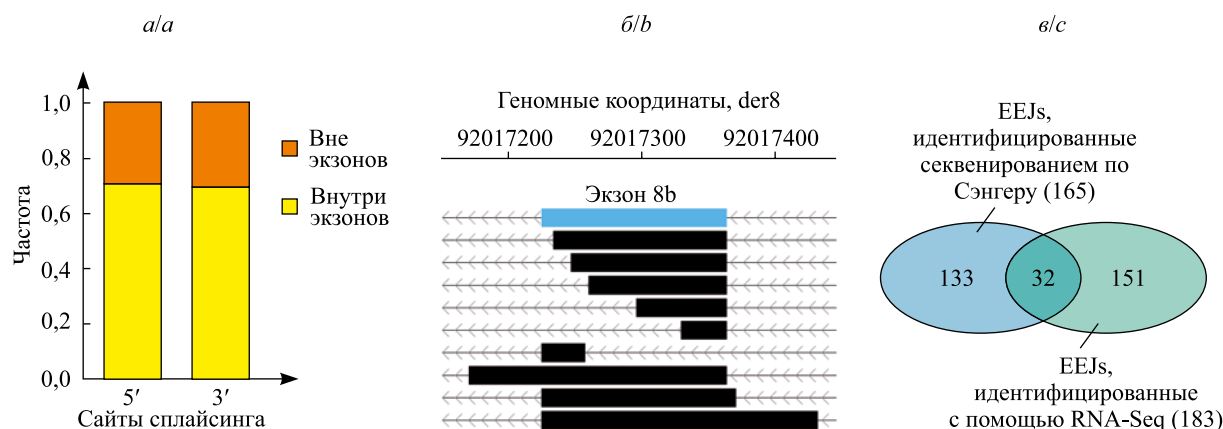


Рис. 1. Набор первичных данных по экспериментальным EEJs гибридного онкогена

*RUNX1-RUNX1T1* может содержать технические артефакты:

- a* – около 70 % сайтов сплайсинга экспериментальных EEJs локализуется внутри канонических экзонов гибридного онкогена *RUNX1-RUNX1T1*;
- б* – канонический экзон 8b (выделен голубым цветом) гибридного онкогена *RUNX1-RUNX1T1* является классическим примером экзона, с которым ассоциировано множество альтернативных 5'- и 3'- сайтов сплайсинга, причем большая часть таких сайтов имеет внутриэкзонную локализацию. Альтернативные экзоны показаны черными прямоугольниками, фланкирующие интронные области изображены горизонтальными линиями, а направление транскрипции указано с помощью стрелок;
- в* – два разных подхода в идентификации EEJs (секвенирование по Сэнгеру и RNA-Seq) дают только частично перекрывающиеся результаты

Fig. 1. Dataset on the experimental EEJs of the fusion oncogene *RUNX1-RUNX1T1* may contain technical artifacts:

- a* – about 70 % of the splice sites of experimental EEJs are localized within the canonical exons of the fusion oncogene *RUNX1-RUNX1T1*;
- b* – the canonical exon 8b (sky blue rectangle) of the fusion oncogene *RUNX1-RUNX1T1* is a classic example of an exon with which many alternative 5' and 3' splice sites are associated.

Herewith, the most of these sites demonstrate intraexon localization. On this picture, alternative exons are shown in black rectangles, flanking intron regions are depicted by horizontal lines, and the direction of transcription is indicated by arrows;  
*c* – two different approaches to identifying EEJs (Sanger sequencing and RNA-Seq) give only partially overlapping results

они практически не попадают в интронные участки. Такой паттерн сплайсинга может свидетельствовать о загрязнении данных техническими артефактами обратной транскрипции, образующимися на основе зрелой, сплайсированной мРНК в результате внутримолекулярной смены матрицы (англ. *intramolecular template switching*, ITS). При ITS обратная транскриптаза отсоединяется от РНК-матрицы и повторно садится на нее же в нижележащем участке, пропустив часть нуклеотидной последовательности. Синтез кДНК при этом не останавливается, из-за чего конечная молекула будет содержать ложные EEJs, не существующие в реальной клетке (англ. *false EEJs*, fEEJs) [22]. Таким образом, для последующей корректной интерпретации наших данных необходимо предварительно оценить, насколько велика в них доля технического шума и артефактов обратной транскрипции.

### Оценка технического шума в экспериментальных данных

Генерация fEEJs за счет ITS осуществляется либо случайным образом, либо закономерно. Первый вариант предполагает, что ITS происходит на произвольном участке РНК в силу стохастических процессов, что подразумевает невозможность воспроизведения fEEJs в повторных экспериментах. Однако наши данные опровергают это предположение. Как видно на рис. 2, а и б, большинство ПЦР-продуктов, которые секвенировались в ходе подготовки группы **кДНК**, устойчиво воспроизводятся даже в условиях ПЦР с повышенной температурой отжига, а также в независимых повторях при клонировании EST в вектор. Тем не менее именно такую картину мы должны наблюдать при ITS, закономерно происходящей на определенных участках РНК.

Единственная экспериментально подтвержденная на сегодняшний день модель, описывающая механизм ITS такого типа, представлена на рис. 2, в, согласно которой признаками потенциальных fEEJs могут служить участки микрогомологии длиной 6–8 нуклеотидов в непосредственной близости от сайтов сплайсинга, а также укладка сегмента РНК между ними в устойчивую вторичную структуру [23]. Эти предположения мы проверили с помощью биоинформатического анализа.

Для каждого из первичных EEJs экстрагировались потенциальные области микрогомологии, прилегающие к сайтам сплайсинга, и для них рассчитывались попарные расстояния Левенштейна (т. е. минимальные количества нуклеотидных замен, позволяющих превратить одну последовательность в другую), распределение которых представлено на рис. 2, г. Для части 6- и 7-нуклеотидных последовательностей расстояние равно нулю, что указывает на их полную идентичность. Однако доля таких последовательностей невелика и уменьшается с ростом длины анализируемых сайтов. Более того, как показано на рис. 2, д и е, парные расстояния Левенштейна и процент идентичности для сайтов первичных EEJs суммарной выборки демонстрируют паттерны распределения, схожие с паттернами у канонических EEJs гибридного онкогена и случайно сгенерированных последовательностей, а также почти полностью совпадают с паттерном распределения у EEJs из базы данных NCBI RefSeq.

Укладка области, находящейся между сайтами микрогомологии, в термодинамически устойчивую структуру типа *ствол – петля* позволяет сайтам микрогомологии приблизиться друг к другу, что значительно облегчает ITS. Мы построили профиль свободной энергии Гиббса вдоль полноразмерной РНК гибридного онкогена, начинающейся с внешнего промотора (включает экзоны 1–4b), и наложили на него распределение 5'- и 3'-сайтов сплайсинга. Результаты представлены на рис. 3, б, из которого видно, что мРНК гибридного онкогена имеет три протяженных участка с низкими значениями энергии Гиббса, потенциально способными формировать стабильные вторичные структуры. Теоретически, если эти участки вовлекаются в ITS, то по их краям должны находиться четкие пики локализации 5'- и 3'-сайтов сплайсинга, однако подобной однозначной картины в наших данных не наблюдается. К тому же множество сайтов сплайсинга локализованы внутри вторичных структур гибридной РНК, для остальных же медиана расстояния до шпильки составляет более 100 нуклеотидов (рис. 3, а и в). На наш взгляд, беспрепятственное протекание ITS на таких больших дистанциях является маловероятным событием.

Кроме того, для группы **полного RNA-Seq** мы оценили покрытие EEJs гибридного онкогена чтениями и сравнили их с покрытием EEJs других генов, а также проанализировали распределение длин интронов (рис. 4, а и б). Подавляющее большинство сплайсируемых интронов, которые депонированы в базе данных RefSeq, имеют длину больше 50 нуклеотидов, а покрытие экзонных стыков выше, чем одно чтение из миллиона (англ. *count per million*, CPM). При фильтрации по указанным критериям в выборке **полного RNA-Seq** сохраняются неканонические EEJs, которые успешно идентифицируются другими средствами. Так, при образовании EEJs 12del из центральной части экзона 12 удаляется экзитрон, соединяя два сегмента на его 5'- и 3'-концах. Помимо **полного RNA-Seq**, этот стык обнаруживается в группе **прицельного RNA-Seq**, а также успешно верифицируется при помощи ОТ-ПЦР (рис. 4, в). Отметим, что для верификации использовалась обратная транскриптаза, устойчивая к ITS [22].

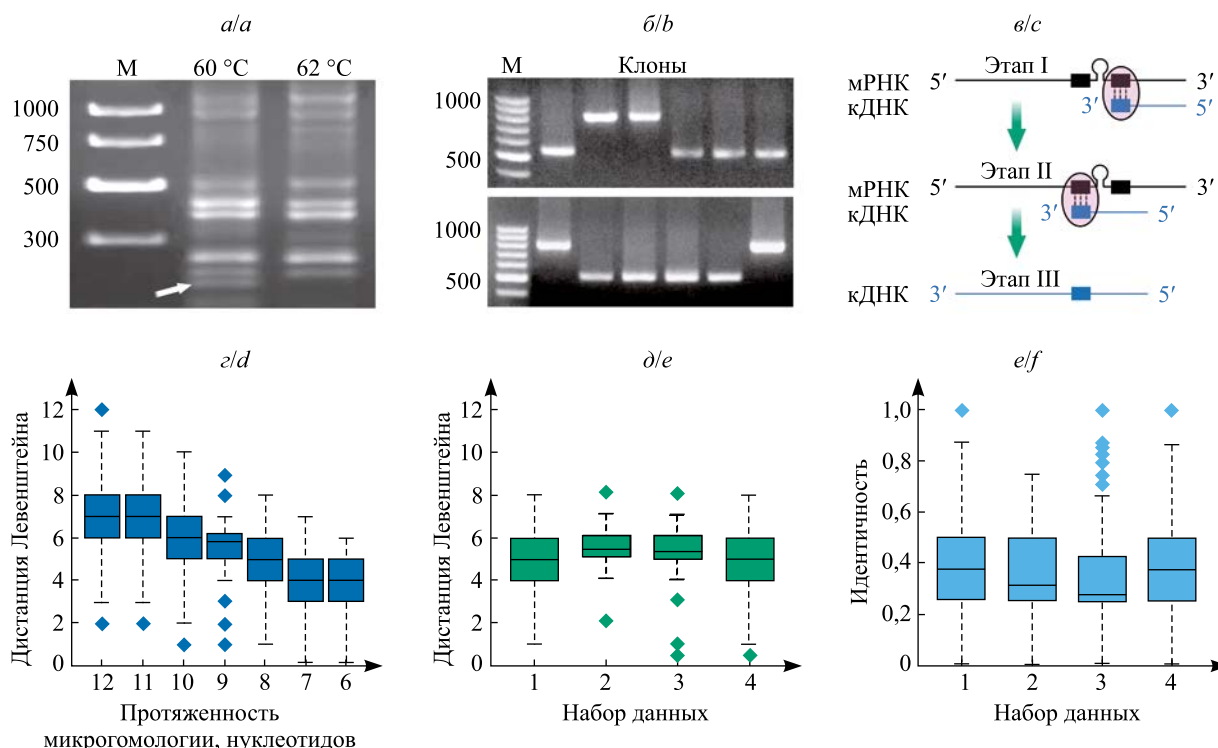


Рис. 2. Идентификация технических артефактов в первичном наборе EEJs гибридного онкогена *RUNX1-RUNX1T1* по сайтам микрогомологии:

- a* – репрезентативный ДНК гель-электрофорез ампликонов кДНК гибридного онкогена *RUNX1-RUNX1T1*, полученных при разных температурах гибридизации прямого и обратного праймеров к экзону 1 гена *RUNX1* и экзону 12а гена *RUNX1T1* соответственно. М – маркеры *Thermo Scientific™ GeneRuler™ Express DNA Ladder*, стрелкой отмечен нестабильный ампликон;
- б* – репрезентативные ДНК гель-электрофорезы результатов скрининга бактериальных клонов-трансформантов из двух независимых серий экспериментов по амплификации и клонированию фрагментов кДНК гибридного онкогена *RUNX1-RUNX1T1*. Ампликоны для клонирования были получены при использовании прямого праймера к экзону 4а гена *RUNX1* и обратного праймера к экзону 12а гена *RUNX1T1*, М – маркеры *Thermo Scientific™ GeneRuler™ 50 bp DNA Ladder*;
- в* – модель внутримолекулярной смены матрицы обратной транскриптазой. Прямоугольниками обозначены области микрогомологии, фланкирующие стабильную вторичную структуру типа *ствол – петля* в мРНК, полупрозрачными овалами показана локализация обратной транскриптазы;
- г* – статистическая характеристика попарных дистанций Левенштейна между последовательностями длиной от 6 до 12 нуклеотидов, фланкирующими 5'- и 3'-сайты сплайсинга всех первичных EEJs;
- д* – статистическая характеристика попарных дистанций Левенштейна для 8-нуклеотидных последовательностей из разных наборов данных: последовательностей, фланкирующих 5'- и 3'-сайты сплайсинга всех первичных EEJs (1), канонических EEJs (2) и NCBI RefSeq EEJs (4), а также случайно сгенерированных 8-нуклеотидных последовательностей (3);
- е* – статистическая характеристика попарной идентичности 8-нуклеотидных последовательностей из наборов данных, описанных в части *д* этого рисунка. Для графиков *г*–*е* горизонтальная линия отражает значение медианы, «усы» – 1,5 межквартильного расстояния

Fig. 2. Microhomology-based identification of technical artifacts in primary set of EEJs of the fusion oncogene *RUNX1-RUNX1T1*:

- a* – representative DNA electrophoregram of amplicons of the cDNA of fusion oncogene *RUNX1-RUNX1T1*. These amplicons were generated at different annealing temperature using forward and reverse primers specific to *RUNX1* exon 1 and *RUNX1T1* exon 12a, respectively. М – markers *Thermo Scientific™ GeneRuler™ Express DNA Ladder*, arrow indicates location of unstable amplicon;
- b* – representative DNA electrophoregrams of the screening results on bacterial transformants with cloned cDNA fragments of the fusion oncogene *RUNX1-RUNX1T1*. Amplicons for cloning were generated in two independent series of experiments using forward and reverse primers specific to *RUNX1* exon 4a and *RUNX1T1* exon 12a, respectively, М – markers *Thermo Scientific™ GeneRuler™ 50 bp DNA Ladder*;
- c* – model of reverse transcriptase intramolecular template switching. Rectangles depict the microhomology sites flanking stable *stem – loop* secondary structure of mRNA, and translucent ovals show the localization of reverse transcriptase;
- d* – statistical summary on pair-wise Levenshtein distances for 6- to 12-mer oligonucleotides flanking 5' and 3' splice sites of all primary EEJs;
- e* – statistical summary on pair-wise Levenshtein distances for 8-mer oligonucleotides flanking 5' and 3' splice sites of all primary EEJs (1), canonical EEJs (2), NCBI RefSeq EEJs (4), or randomly generated 8-mer oligonucleotides (3);
- f* – statistical summary on identity of 8-mer oligonucleotides from datasets described in *e* panel.
- For plots *d*–*f* median is shown by horizontal line, 1.5 IQR is shown by whiskers

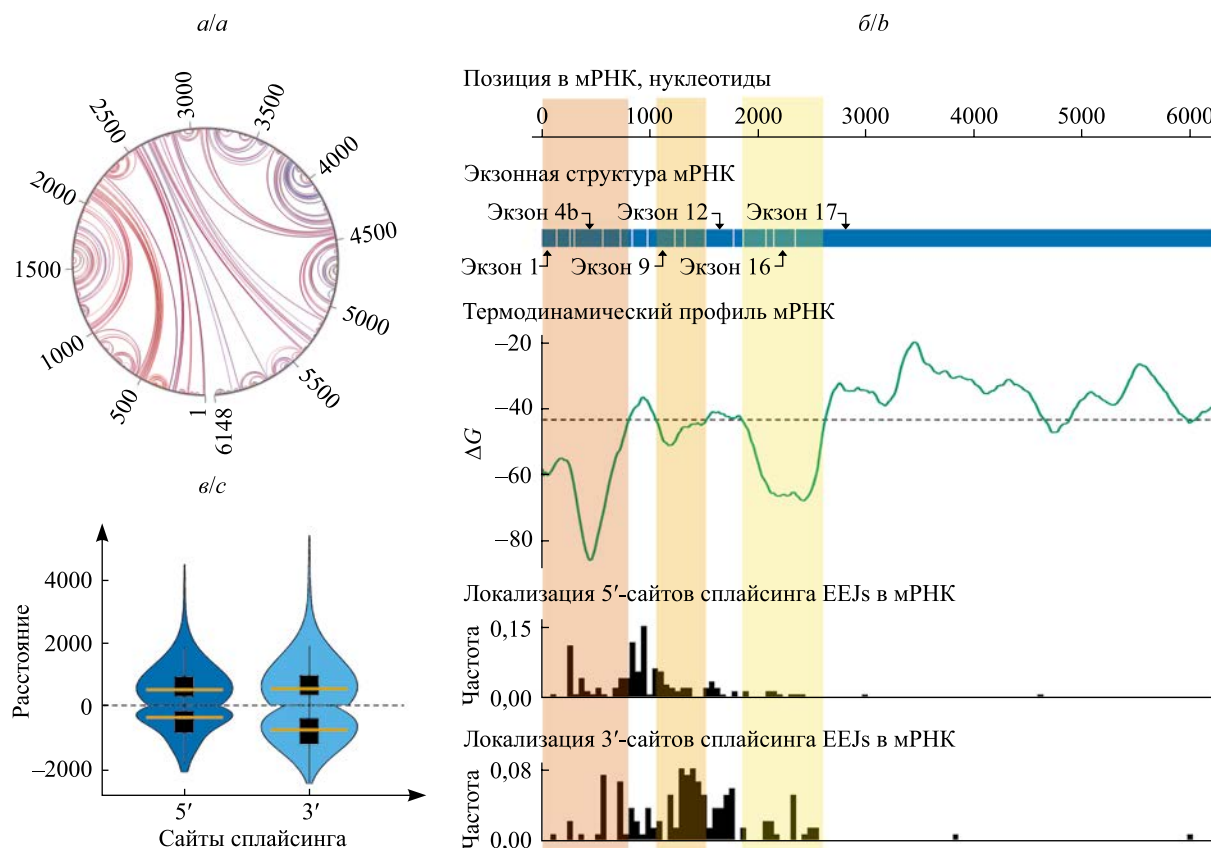


Рис. 3. Термодинамические особенности мажорной изоформы мРНК гибридного онкогена *RUNX1-RUNX1T1*, начинающейся с экзона 1:

*а* – круговая диаграмма вторичной структуры мРНК. Вторичная структура молекулы РНК была рассчитана с помощью приложения Mfold, нуклеотиды пронумерованы по часовой стрелке, красные дуги объединяют нуклеотиды *G* и *C*, синие дуги указывают на пары *A–U*, зеленые дуги – на пары *G–U*;

*б* – соотношение между термодинамическим профилем мРНК и локализацией в ней экзонов, а также сайтов сплайсинга из первичного набора EEJs. На диаграмме экзонной структуры мРНК позиции маркерных экзонов отмечены стрелками, вертикальными широкими полосами выделены три наиболее структурированные области мРНК, для которых значения  $\Delta G$  всегда получаются ниже усредненного значения (горизонтальная линия) по всей молекуле мРНК;

*в* – статистическая характеристика близости (выраженной в количестве нуклеотидов) 5'- и 3'-сайтов сплайсинга всех первичных EEJs, картированных по мРНК, от границ наиболее устойчивых вторичных структур в этой же молекуле РНК.

Для всех 5'-сайтов сплайсинга рассчитывалась их близость к началу наиболее устойчивых вторичных структур, а для 3'-сайтов сплайсинга – к окончанию таких структур. Дальнейший статистический анализ полученных распределений проводился отдельно для 5'- и 3'-сайтов сплайсинга, а также отдельно для тех случаев, когда сайт располагался перед соответствующей координатой вторичной структуры (расстояние со знаком «минус») или после нее (расстояние со знаком «плюс»).

Горизонтальная линия отображает значение медианы, «усы» – 1,5 межквартильного расстояния

Fig. 3. Thermodynamic properties of exon 1 started major mRNA isoform of the fusion oncogene *RUNX1-RUNX1T1*:

*a* – circular diagram of the secondary structure of mRNA. The secondary structure of RNA molecule was calculated using Mfold application, nucleotides are numbered clockwise, *G–C* arcs are drawn in red, *A–U* arcs are shown in blue, and *G–U* arcs are depicted in green;

*b* – the relationship between the thermodynamic profile of mRNA and the localization of exons and splice sites from the primary EEJs in this RNA isoform. On the diagram of the exon structure of mRNA, the positions of some exons are marked by arrows. Three the most structured regions of the mRNA are shaded using vertical wide bands. For these regions,  $\Delta G$  values are always lower than the average value (horizontal line) over the entire mRNA molecule;

*c* – statistical summary on proximity (as number of nucleotides) of 5' and 3' splice sites of all primary EEJs mapped against mRNA to boundaries of the most stable secondary structures in this RNA isoform.

For all 5' and 3' splice sites, their proximity to the beginning and the end of the most stable secondary structures was calculated, respectively. Further statistical analysis of the distance distributions was carried out separately for the 5' and 3' splice sites, and also separately for those cases where the site was located upstream (the distance with a minus sign) or downstream (the distance with a plus sign) the corresponding coordinate of the secondary structure.

Median is shown by horizontal line, 1.5 IQR is shown by whiskers

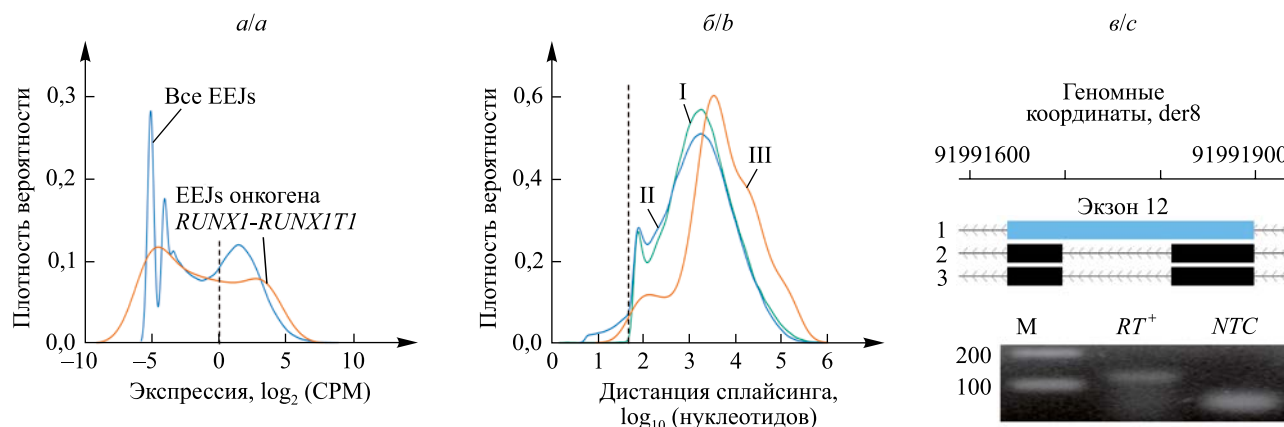


Рис. 4. Поиск технических артефактов в наборе EEJs гибридного онкогена *RUNX1-RUNXIT1*, идентифицированных с помощью RNA-Seq:

*a* – существенная часть EEJs, идентифицированных с помощью технологии RNA-Seq в транскриптоме клеток линии Kasumi-1 как для гибридного онкогена *RUNX1-RUNXIT1*, так и для других генов, характеризуются очень низким уровнем экспрессии. Для снижения шума в экспериментальных данных такие плохо подтверждаемые EEJs могут быть устранены на этапе разведочного анализа данных;

*б* – распределение EEJs по дистанциям сплайсинга (длине интронов).

На графике представлены результаты анализа трех наборов данных:

EEJs, идентифицированных в тех молекулах РНК человека, которые депонированы в базе данных NCBI RefSeq (I);

всех EEJs, идентифицированных с помощью технологии RNA-Seq в транскриптоме клеток линии Kasumi-1 (II);

EEJs гибридного онкогена *RUNX1-RUNXIT1*, идентифицированных с помощью технологии RNA-Seq в транскриптоме клеток линии Kasumi-1 (III);

*в* – EEJs гибридного онкогена *RUNX1-RUNXIT1*, идентифицированные по технологии RNA-Seq, могут быть успешно валидированы методом обратной транскрипции и последующей амплификации с помощью полимеразной цепной реакции. В качестве примера представлена валидация экзитрона, обнаруженного в каноническом экзоне 12 (трек 1, выделен голубым цветом) гибридного онкогена *RUNX1-RUNXIT1* методом тотальной (трек 2) или направленной (трек 3) RNA-Seq.

На электрофореграмме показан результат контрольной (без кДНК-матрицы, NTC) и специфической (после обратной транскрипции, RT<sup>+</sup>) амплификации.

M – маркеры Thermo Scientific™ GeneRuler™ DNA Ladder Mix

Fig. 4. Identification of technical artifacts among EEJs of the fusion oncogene *RUNX1-RUNXIT1* detected by RNA-Seq:

*a* – substantial part of EEJs detected by RNA-Seq in the transcriptome of Kasumi-1 cells demonstrates very low level of expression. This remains true for both fusion oncogene EEJs and EEJs of other genes expressed in leukemia cells.

To reduce the noise in the experimental data, such poorly supported EEJs can be eliminated during the exploratory data analysis phase; *b* – distribution of EEJs on splicing distances (or introns length).

This plot shows the results of the analysis of three data sets: EEJs retrieved from human NCBI RefSeq RNAs (I);

all EEJs identified by RNA-Seq in the transcriptome of Kasumi-1 cells (II);

EEJs of the fusion oncogene *RUNX1-RUNXIT1* detected by RNA-Seq in the transcriptome of Kasumi-1 cells (III);

*c* – RNA-Seq detected EEJs of the fusion oncogene *RUNX1-RUNXIT1* can be successfully validated by reverse transcription and polymerase chain reaction.

Validation of exitron located in canonical exon 12 (track 1, sky blue rectangle) of the fusion oncogene *RUNX1-RUNXIT1* is presented as example. Exitron was detected by RNA-Seq on total RNA (track 2) or only RNA

of the fusion oncogene (track 3). Electrophoregram demonstrates the results of control (without cDNA template, NTC) and specific (after reverse transcription, RT<sup>+</sup>) amplification. M – markers Thermo Scientific™ GeneRuler™ DNA Ladder Mix

Таким образом, абсолютное большинство EEJs не демонстрируют признаков вовлеченности в ITS, что позволяет рассматривать их как истинные. Тем не менее для повышения надежности дальнейшего анализа мы провели фильтрацию потенциальных fEEJs по следующим параметрам: 1) дистанция сплайсинга – не менее 50 нуклеотидов; 2) EEJs картируются по минус-цепи; 3) сайты сплайсинга не фланкированы участками микрогомологии (для последовательностей длиной 6–8 или 9–12 нуклеотидов расстояние Левенштейна больше 0 или 1 соответственно); 4) расстояние от 3'-сайта сплайсинга до 3'-конца термодинамически устойчивой структуры в канонической мРНК не превышает 12 нуклеотидов; 5) для EEJs *полного RNA-Seq* покрытие чтением – не ниже 1 CPM. Такая фильтрация исключает из первоначального списка 36,4 % EEJs и сокращает их количество до 201. Именно эти EEJs, а также образуемые ими экзоны анализировались нами в дальнейшем.

### Идентификация основных мод альтернативного сплайсинга пре-мРНК гибридного онкогена *RUNX1-RUNXIT1*

Мы классифицировали сплайсинговые события, которые обнаруживаются внутри мРНК *RUNX1-RUNXIT1*, по нескольким модам, или типам, альтернативного сплайсинга (рис. 5). Как можно видеть,

и для группы *кДНК*, и для группы *полного RNA-Seq* паттерн распределения типов альтернативного сплайсинга оказывается схожим. Примерно в 80 % случаев EEJs участвуют в одном из трех типов сплайсинга (использование альтернативного 5'- или 3'-сайта сплайсинга либо пропуск нескольких каскадных экзонов подряд). Однако картина значительно меняется, если для EEJs *полного RNA-Seq* проанализировать количество подтверждающих их чтений. Так, хотя EEJs, использующие альтернативный вариант 5'-сайта сплайсинга, составляют 30,6 % уникальных сплайсинговых событий, количественно на них приходится лишь 3,7 % чтений. Максимальным же уровнем покрытия обладают пропуски одиночных каскадных экзонов и альтернативные 5'-концевые экзоны – суммарно на них приходится более 80 % подтверждающих чтений (рис. 5, в).

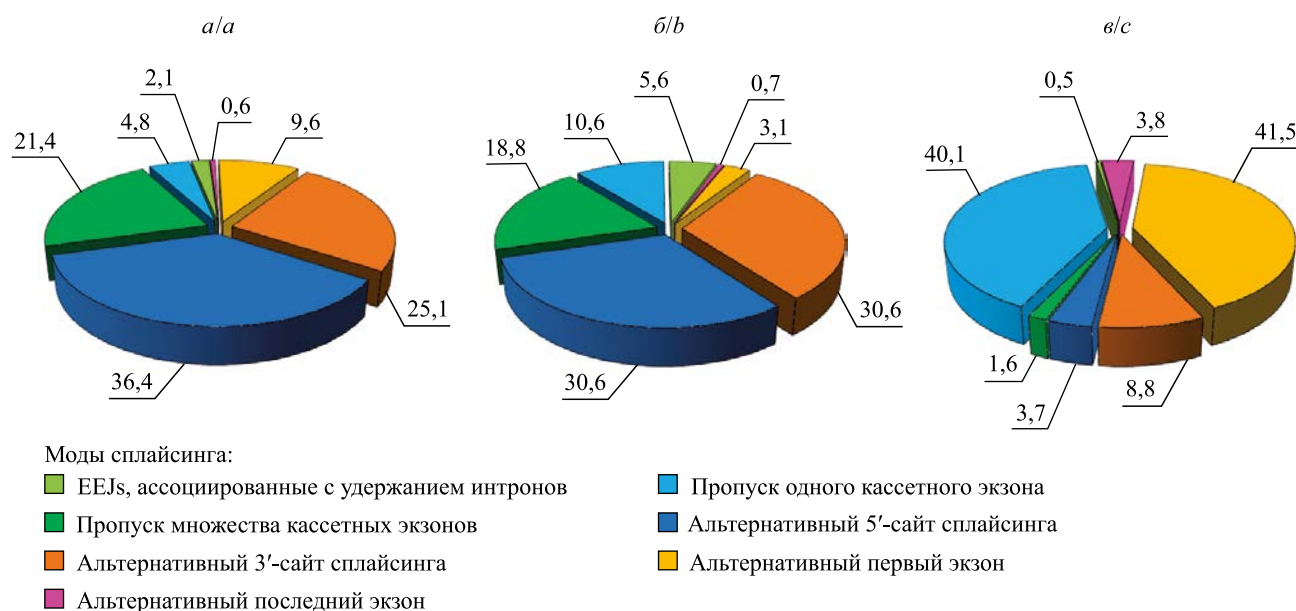


Рис. 5. Классификация EEJs гибридного онкогена *RUNX1-RUNX1T1* по модам альтернативного сплайсинга.

Представлены частоты встречаемости (в процентах) семи основных мод сплайсинга среди EEJs, идентифицированных секвенированием по Сэнгеру (а), а также обнаруженных методом RNA-Seq без учета (б) или с учетом (в) уровня экспрессии

Fig. 5. Classification of EEJs of the fusion oncogene *RUNX1-RUNX1T1* according to modes, or types, of alternative splicing.

The figure shows the frequency (in percent) of the seven main splicing modes among EEJs identified by Sanger sequencing (a) and by RNA-Seq without (b) or with taking into account the abundance of junctions (c)

Такое распределение соотносится с нашей оценкой экспрессии полноразмерных мРНК *RUNX1-RUNX1T1* при помощи *Salmon* [14]. При этом основная часть транскриптов гибридного онкогена – кодирующие полноразмерный либо же усеченный по С-концу белок транскрипты, состоящие из стандартных экзонов гибридного онкогена, а также канонических каскадных экзонов 12а, 15а и 17а. Эти транскрипты нарабатываются с использованием различных промоторных областей [24]. Доля же мРНК с неканоническими вариантами экзонов не превышает нескольких процентов от общего пула транскриптов *RUNX1-RUNX1T1*. Именно эти варианты РНК являются источниками альтернативных 5'- и 3'-сайтов сплайсинга.

Для EEJs 12a12, 15a15, 17a17 и 12del, которые присутствуют в сборке *полного RNA-Seq*, мы оценили уровни экспрессии по отношению к стыку между экзонами 6 и 8b гибридного онкогена. Судя по всему, данный стык присутствует в большинстве гибридных транскриптов, служит надежным маркером РНК *RUNX1-RUNX1T1* в клетке и может использоваться для определения примерной доли того или иного сплайсингового события в общем пуле транскриптов гибридного онкогена. При оценке использовались три разных способа: 1) определение доли включающих данный вариант сплайсинга мРНК по результатам *Salmon*; 2) вычисление отношения числа чтений, подтверждающих экспериментальный стык, к числу чтений, подтверждающих стык 6-8b; 3) применение количественной ПЦР с использованием устойчивой к ITS обратной транскриптазы. Результаты представлены на рис. 6.

Как можно видеть, результаты прямого подсчета чтений и количественной ПЦР хорошо соотносятся друг с другом. Несмотря на то что оценка *Salmon* имеет тенденцию к завышению долей отдельных

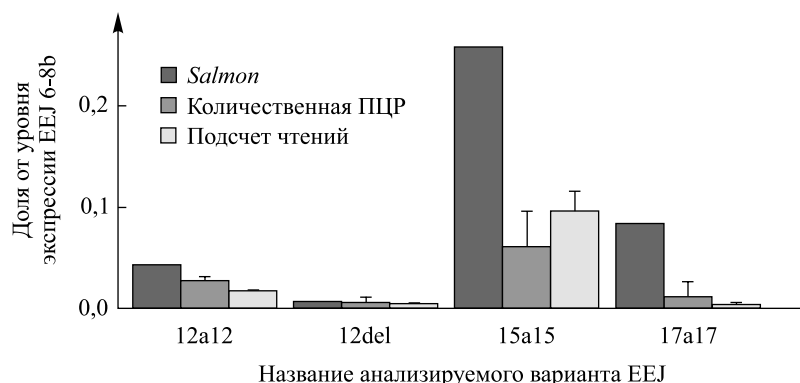


Рис. 6. Кросс-валидация уровней экспрессии EEJs разными методами.

Показаны средние значения уровня экспрессии EEJs по отношению к EEJ 6-8b и их стандартные отклонения

Fig. 6. Cross-validation of EEJs expression using different approaches.

This plot shows the results from three independent experiments (as arithmetic mean plus/minus standard deviation).

The expression of each EEJ of interest was normalized relative to EEJ 6-8b abundance

сплайсинговых событий, в целом все три метода оценивают соотношение уровней экспрессии EEJs схожим образом. Это позволяет нам выделить три категории сплайсинговых событий в мРНК гибридного онкогена:

- мажорные события. К ним можно отнести все канонические EEJs, входящие в состав полноразмерных транскриптов, в том числе стык 16-17a и EEJs, генерируемые при транскрипции с внешнего промотора;
- умеренные события, к которым относится в первую очередь использование канонических кассетных экзонов 12a и 15a, а также некоторые неканонические события: удлинение экзона 1, использование стыка RUNX1/PR07-4b (см. ниже). Такие EEJs встречаются в 1–10 % транскриптов гибридного онкогена (при оценке количественной ПЦР или подсчете чтений);
- минорные события. Их доля колеблется около 1 % и ниже либо имеет стандартное отклонение, близкое к значению средней доли. В эту группу попадают EEJs 17a17, 12del, а также абсолютное большинство EEJs на основе альтернативных вариантов экзонов из групп *кДНК* и *прицельного RNA-Seq*. Именно экстремально низкий уровень экспрессии объясняет отсутствие большинства таких EEJs в сборке мРНК из полноразмерной библиотеки чтений.

### Идентификация «горячих областей» альтернативного сплайсинга пре-мРНК гибридного онкогена *RUNX1-RUNX1T1*

EEJs из нашего набора данных образуют 332 уникальных экзона, имеющих длину не менее 25 нуклеотидов и лежащих на минус-цепи ДНК. Их фильтрация по описанным ранее критериям сокращает количество экзонов до 228, из которых 179 перекрываются с каноническими экзонными областями. Из оставшихся 49 экзонов большая часть лежит в области гибридного интрона между экзонами 6 и 8b. Подавляющее их большинство описано в работе [2]. Это единственный участок гибридного онкогена, на котором наблюдается экзонизация множества внутренних областей интрона. Помимо них, данная подгруппа областей включает кассетные экзоны 12a, 15a и 17a (экзон 12a имеет два экзонных варианта, 15a – один, 17a – три), а также 91-нуклеотидную вставку интрона между экзонами 10 и 11. Отметим, что уровни фильтрации, выбранные нами, оказались достаточно строгими и отсекали небольшое количество верифицированных экспериментально EEJs, к примеру канонический стык 16-17.

Картирование экзонов гибридного онкогена на искусственную мета-мРНК, содержащую все канонические экзоны, позволяет заключить, что примерно 80 % альтернативных экзонных вариантов генерируются в «горячей области», включающей экзоны 4a/b, 6, 8b, 9, 11 и 12. Экзоны 4a и 12 демонстрируют высокую частоту экзитроноподобного сплайсинга, в результате чего теряется центральная часть канонической области, а 5'- и 3'-сегменты становятся самостоятельными экзонами. Другие экзоны генерируют варианты преимущественно за счет усечения по 5'- и 3'-концам, а также экзонизации прилегающих участков. Все эти экзонные варианты относятся к минорной фракции. Как видно из рис. 7, схожее распределение числа экзонных вариантов наблюдается внутри всех независимых групп. Хотя в группах *кДНК* и *прицельного RNA-Seq* присутствует незначительное количество экзонных вариан-

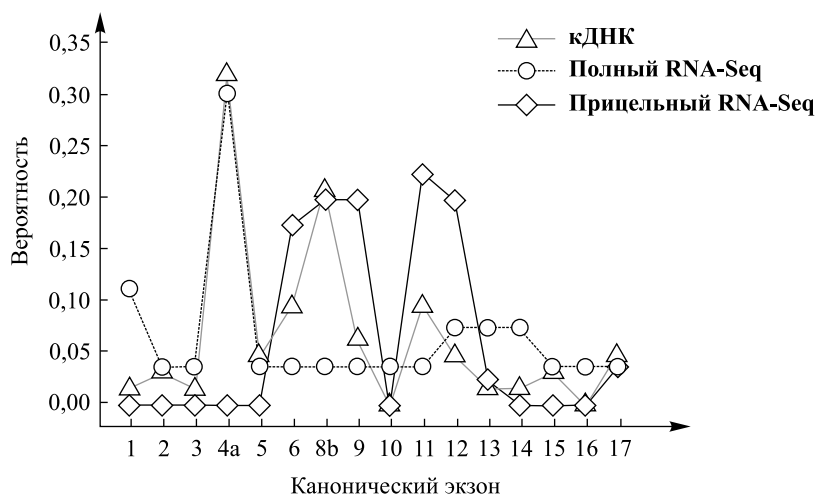


Рис. 7. Разнообразие альтернативных вариантов экзонов, генерируемых на основе канонических экзонов гибридного онкогена *RUNX1-RUNX1T1*. Представлено распределение частот встречаемости генерируемых вариантов среди всех вариантов экзонов гибридного онкогена *RUNX1-RUNX1T1*, обнаруживаемых с помощью указанного метода

Fig. 7. A diversity of alternative exons generated on the basis of the canonical exons of the fusion oncogene *RUNX1-RUNX1T1*. This plot shows the frequency distribution of the occurrence of variants generated on the basis of given canonical exon among all variants of exons of the fusion oncogene *RUNX1-RUNX1T1* detected by indicated method

тов, образующихся искусственно, через отсечку фрагмента экзона ПЦР-праймером, это не приводит к исчезновению «горячих областей» (данные не показаны). Указанные наблюдения подтверждают правильность наших выводов.

### Вклад альтернативных промоторов и сайтов полиаденилирования в разнообразие альтернативных транскриптов гибридного онкогена *RUNX1-RUNX1T1*

Известно, что транскрипция гибридного онкогена обеспечивается главным образом благодаря двум каноническим промоторам. РНК, синтезированные с внешнего промотора, могут включать экзоны 1 и 2, а также 3, который сплайсируется с укороченным вариантом экзона 4-4b. GC-богатый внутренний промотор является источником транскриптов, имеющих экзон 4a в качестве 5'-концевого. Также есть свидетельства наличия внутренних промоторных областей вблизи экзона 8b. Однако анализ данных *полного RNA-Seq* обнаруживает промоторную активность в еще одной области – на участке ранее предсказанного нами седьмого промоторного региона [4]. Ее использование приводит к образованию 5'-концевого экзона с координатами der8:92068127–92068487, сплайсирующегося с областью 4b (рис. 8, а).

Указанный экзон отстоит от канонической точки начала транскрипции примерно на 1 т. п. о. и также лежит в GC-богатой области. Прямой подсчет чтений оценивает долю включающих его транскриптов на уровне  $2,7 \pm 1,5$  % от стыка 6-8b. Его присутствие подтверждается ОТ-ПЦР [14], а также предсказаниями NCBI-Gnomon (XM\_005261068.3). Стоит отметить, что транскрипты, содержащие экзон PR07, лишены канонических точек инициации трансляции, локализованных внутри экзонов 3 и 4a.

База данных GENCODE v.27 содержит координаты лишь одного сайта полиаденилирования для интактного гена *RUNX1T1*. Эта точка с хромосомной координатой der8:91956502 располагается в ассоциированной с экзоном 17 3'-нетранслируемой области (3'-НТО) и генерирует концевой экзон длиной около 4 т. п. о. Координаты сайта полиаденилирования экзона 17a в базах данных отсутствуют, однако можно опереться на результаты авторов [25], впервые описавших указанный экзон. Согласно этим данным, сайт полиаденилирования находится в положении der8:91964725, что дает концевой участок длиной 1650 нуклеотидов. Для того чтобы подтвердить использование этих сайтов полиаденилирования и выявить новые, мы проанализировали библиотеку чтений *прицельного RNA-Seq* при помощи программы *KLEAT-2.0* [9]. С ее помощью удалось идентифицировать два сайта полиаденилирования внутри экзонов 17 и 17a (рис. 8, б).

Оба сайта находятся значительно ближе к 5'-концу экзона, чем канонические сайты, что приводит к образованию укороченных 3'-НТО. Для экзона 17 координата точки полиаденилирования der8:91959537, длина 3'-НТО – порядка 850 нуклеотидов. Для экзона 17a полиаденилирование происходит в точке

der8:91966220, длина 3'-НТО составляет всего лишь 100 нуклеотидов. При этом реконструированные с помощью *Cufflinks* транскрипты содержат несколько вариантов объединенного с 3'-НТО экзона 17, самый длинный из которых включает 5473 нуклеотида и захватывает канонический сайт полиаденилирования. Как можно видеть, большая часть этого сегмента мРНК имеет высокое покрытие чтениями, которое, однако, в области канонического сайта полиаденилирования *RUNX1T1* резко падает. В отношении сайта, предложенного в [25], для экзона 17а такого падения не наблюдается, однако для этой области *Cufflinks* также реконструирует длинные, хорошо покрытые 3'-НТО.

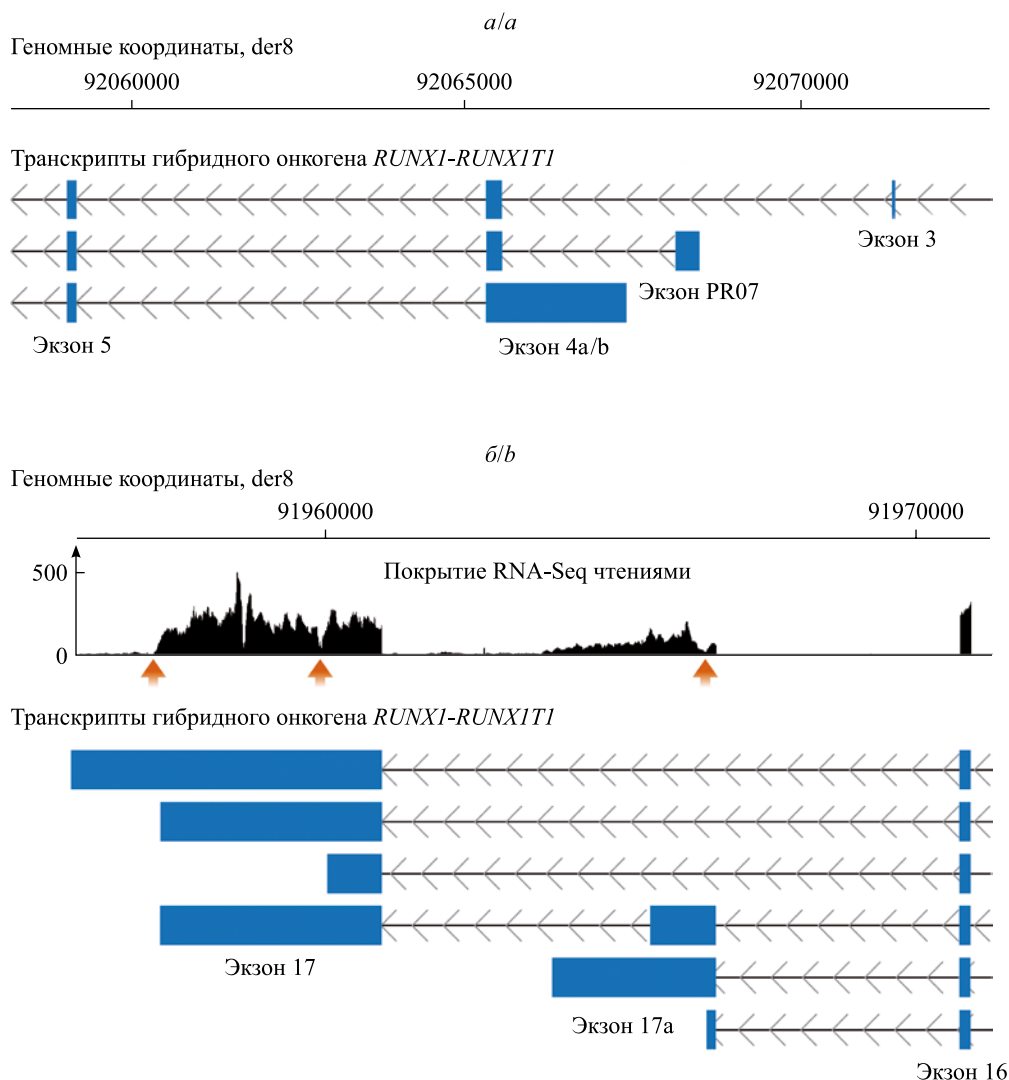


Рис. 8. Вклад альтернативных промоторных областей *RUNX1*-локуса и сайтов альтернативного полиаденилирования в разнообразие мРНК гибридного онкогена *RUNX1-RUNX1T1*:  
а – вклад внешнего, внутреннего промоторов и промотора *RUNX1/PR07* в формирование 5'-конца мРНК гибридного онкогена *RUNX1-RUNX1T1*;  
б – вклад точек альтернативного полиаденилирования в формирование 3'-конца мРНК гибридного онкогена *RUNX1-RUNX1T1*.

Локализация идентифицированных с применением алгоритма *KLEAT-2.0* и анализа покрытия сайтов полиаденилирования показана с помощью вертикальных стрелок. Экзоны обозначены прямоугольниками, интроны – горизонтальными линиями, направление транскрипции – стрелками в «теле» транскриптов

Fig. 8. The contribution of alternative promoter regions of the *RUNX1* locus and alternative polyadenylation sites to the mRNA diversity of the fusion oncogene *RUNX1-RUNX1T1*:

а – the contribution of the external, internal promoters and the *RUNX1/PR07* promoter to the formation of the 5'-end of the fusion oncogene *RUNX1-RUNX1T1* mRNA;

б – the contribution of alternative polyadenylation sites to the formation of the 3'-terminus of fusion oncogene *RUNX1-RUNX1T1* mRNA. Localization of the polyadenylation sites identified by the algorithm *KLEAT-2.0* and analyzing the coverage is shown using the vertical arrows.

The exons are indicated by rectangles, the introns are indicated by horizontal lines, the direction of transcription is indicated by arrows in the «body» of the transcripts

Вышесказанное свидетельствует о том, что для мРНК, заканчивающихся экзоном 17 (а может быть, и экзоном 17а), функционируют несколько сайтов полиаденилирования, за счет чего в клетке коэкспрессируются изоформы мРНК как с полноразмерными, так и укороченными 3'-НТО.

### Заключение

Огромное разнообразие вариантов мРНК *RUNX1-RUNXIT1* ставит вопрос о причинах такого разнообразия, а также о закономерностях, которым подчиняется генерация транскриптов этого гена. В ходе исследования мы получили большой массив данных из различных источников, опираясь на который оценили пул альтернативных экзонов гибридного онкогена как с количественной, так и с качественной стороны. Это дало возможность сформулировать некоторые общие принципы, на основе которых генерируется пул мРНК гибридного онкогена.

Количественный анализ позволяет разделить сплайсинговые события в мРНК *RUNX1-RUNXIT1* на три категории. К фракциям мажорных и умеренных событий относятся преимущественно канонические экзоны и EEJs, в то время как большинство альтернативных экзонных вариантов попадает в минорную фракцию. Абсолютное большинство альтернативных экзонных вариантов образуется на основе какого-либо канонического экзона и сохраняет неизменной хотя бы одну из его внешних границ. Этим объясняется тот факт, что приблизительно для 45 % 5'- и 3'-сайтов сплайсинг происходит по каноническим нуклеотидным парам GC-AG и GT-AG. Для остальных сайтов ярко выраженные предпочтения в составе динуклеотидов отсутствуют. Из этих наблюдений можно заключить, что распознавание неканонических сайтов сплайсинга происходит в результате стохастических процессов, сопровождающих сборку сплайсосомы. Поскольку для образования экзонов с модификацией сразу по двум концам должны произойти два независимых друг от друга случайных события, их количество в выборке невелико. Таким образом, минорная фракция экзонов является результатом биологического шума [26]. Однако при этом минорные экзонные варианты устойчиво воспроизводятся от эксперимента к эксперименту, что не позволяет рассматривать их как полностью случайные события. Вероятно, их распознавание определяется локальными особенностями структуры премРНК, которые также определяют и существование «горячих областей» альтернативного сплайсинга.

Суммарную долю транскриптов, включающих экзоны минорной фракции, сложно оценить. Во-первых, их максимальное число ограничено лишь количеством потенциальных сайтов сплайсинга в первичной мРНК. Во-вторых, точные оценки долей затруднены ограничениями существующих количественных методик, а также нестабильной экспрессией экзонных вариантов. В-третьих, одна мРНК может включать несколько минорных сплайсинговых событий. Однако, основываясь на кросс-валидированных RNA-Seq данных и оценках путем количественной ПЦР, можно предположить, что верхняя граница доли транскриптов минорной фракции в общем пуле не превышает нескольких процентов. Наиболее эффективными из опробованных нами методов изучения их структуры стоит признать секвенирование по Сэнгеру и цельный RNA-Seq.

На основании результатов 3'-RACE и *полного RNA-Seq* мы предполагаем коэкспрессию в пуле мРНК *RUNX1-RUNXIT1* 3'-НТО различной длины. Поддержание определенного соотношения между ними может иметь важное физиологическое значение. Известно, что варьирование длины 3'-НТО мРНК является распространенным регуляторным механизмом. В частности, активированные Т-лимфоциты в ходе пролиферации экспрессируют мРНК с укороченными 3'-НТО, что способствует синтезу большого количества белка [27]. Уровень экспрессии транскриптов гибридного онкогена в t(8;21)-положительных лейкозных бластах значительно выше, чем в дифференцированных моноцитах и гранулоцитах с этой мутацией [28]. Данное увеличение коррелирует с экспрессией укороченных вариантов 3'-НТО, а синтез белка на таких транскриптах повышается до 5 раз [28]. Также мРНК с длинными 3'-концевыми участками могут выступать в качестве ловушек миРНК, опосредованно регулируя экспрессию других генов [29].

Использование неканонических промоторов может вносить вклад в разнообразие белков, которые синтезируются с гибридного локуса. Помимо промоторных областей, локализованных в области гибридного интрона, область RUNX1/PR07 также может генерировать транскрипты. Биоинформатический анализ показывает, что такие мРНК лишены канонических точек начала трансляции в экзонах 3 и 4а. Если трансляция на них может инициироваться в других областях, к примеру на участке 8а/б, это приведет к наработке интактных вариантов белка RUNX1T1, лишённого RUNX1-части. Примечательно, что именно такие, укороченные, изоформы белка можно обнаружить при помощи вестерн-блота в t(8;21)-положительных клетках [4].

Таким образом, все три основных механизма генерации альтернативных мРНК – альтернативные промоторы, альтернативный сплайсинг и альтернативные сайты полиаденилирования – вносят существенный вклад в разнообразие альтернативных транскриптов гибридного онкогена *RUNX1-RUNXIT1*. Вероятно, они играют важную роль в поддержании жизнеспособности раковой клетки и лейкозогенеза; в то же время особенности минорной фракции экзонов позволяют наблюдать общие принципы

функционирования систем сплайсинга. Более глубокое изучение этих принципов, их реализации в клетках других типов, а также функциональных ролей 3'-НТО и неканонических промоторов является предметом будущих исследований в этой области.

### Библиографические ссылки

1. Peterson LF, Boyapati A, Ahn E, Biggs JR, Okumura AJ, Lo M-C, et al. Acute myeloid leukemia with the 8q22;21q22 translocation: secondary mutational events and alternative t(8;21) transcripts. *Blood*. 2007;110(3):799–805. DOI: 10.1182/blood-2006-11-019265.
2. LaFiura KM, Edwards H, Taub JW, Matherly LH, Fontana JA, Mohamed AN, et al. Identification and characterization of novel AML1-ETO fusion transcripts in pediatric t(8;21) acute myeloid leukemia: a report from the Children's Oncology Group. *Oncogene*. 2008;27(36):4933–4942. DOI: 10.1038/onc.2008.134.
3. Grinev VV, Migas AA, Kirsanova AD, Mishkova OA, Siomava N, Ramanouskaya TV, et al. Decoding of exon splicing patterns in the human *RUNX1-RUNX1T1* fusion gene. *International Journal of Biochemistry and Cell Biology*. 2015;68:48–58. DOI: 10.1016/j.biocel.2015.08.017.
4. Migas AA, Mishkova OA, Ramanouskaya TV, Ilyushonak IM, Aleinikova OV, Grinev VV. *RUNX1T1/MTG8/ETO* gene expression status in human t(8;21)(q22;q22)-positive acute myeloid leukemia cells. *Leukemia Research*. 2014;38(9):1102–1110. DOI: 10.1016/j.leukres.2014.06.002.
5. Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research*. 2001;29(9):e45. DOI: 10.1093/nar/29.9.e45.
6. Ramakers C, Ruijter JM, Deprez RHL, Moorman AF. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neuroscience Letters*. 2003;339(1):62–66. DOI: 10.1016/S0304-3940(02)01423-4.
7. Ruijter JM, Ramakers C, Hoogaars WM, Karlen Y, Bakker O, van den Hoff MJ, et al. Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Research*. 2009;37(6):e45. DOI: 10.1093/nar/gkp045.
8. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–2120. DOI: 10.1093/bioinformatics/btu170.
9. Birol I, Raymond A, Chiu R, Nip KM, Jackman SD, Kreitzman M, et al. Kleat: cleavage site analysis of transcriptomes. In: Altman RB, Dunker AK, Hunter L, Ritchie MD, Murray TA, Klein TE, editors. *Biocomputing-2015. Pacific Symposium on Biocomputing; 2015 January 4–8; Fairmont Orchid, Hawaii*. [S. l.]: [s. n.]; 2015. p. 347–358. DOI: 10.1142/9789814644730\_0034.
10. Xiao Z, Greaves MF, Boffler P, Smith M. Molecular characterization of genomic AML1-ETO fusions in childhood leukemia. *Leukemia*. 2001;15(12):1906–1913. DOI: 10.1038/sj.leu.2402318.
11. Zhang Y, Strissel P, Strick R, Chen J, Nucifora G, Le Beau MM, et al. Genomic DNA breakpoints in AML1/RUNX1 and ETO cluster with topoisomerase II DNA cleavage and DNase I hypersensitive sites in t(8;21) leukemia. *Proceedings of the National Academy of Sciences*. 2002;99(5):3070–3075. DOI: 10.1073/pnas.042702899.
12. Kent WJ. BLAT – the BLAST-like alignment tool. *Genome Research*. 2002;12(4):656–664. DOI: 10.1101/gr.229202.
13. Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, et al. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res*. 2016;44(D1):D717–725. DOI: 10.1093/nar/gkv1275.
14. Ильюшёнко ИИ, Гунько ЕП, Антонович МЛ, Яцков НН, Кустанович АМ, Сухаревский АЮ и др. Изучение закономерностей сплайсинга РНК гибридного онкогена *RUNX1-RUNX1T1* человека с помощью методов интеллектуального анализа данных и высокопроизводительного секвенирования. В: Кильчевский АВ, главный редактор. *Молекулярная и прикладная генетика. Сборник научных трудов. Том 23*. Минск: Институт генетики и цитологии НАН Беларуси; 2017. с. 92–101.
15. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nature Biotechnology*. 2011;29(1):24–26. DOI: 10.1038/nbt.1754.
16. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*. 2017;14(4):417–419. DOI: 10.1038/nmeth.4197.
17. Pages H, Aboyoun P, Gentleman R, DebRoy S. Biostings: Efficient manipulation of biological strings. R package version 2.30.1. 2014. DOI: 10.18129/B9.bioc.Biostings.
18. Morgan M, Pages H, Obenchain V, Hayden N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 1.34.1. 2019. DOI: 10.18129/B9.bioc.Rsamtools.
19. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*. 2013;9(8):e1003118. DOI: 10.1371/journal.pcbi.1003118.
20. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*. 2011;6:26. DOI: 10.1186/1748-7188-6-26.
21. Grinev VV, Ilyushonak IM, Clough R, Nakjang S, Smink J, Martinez-Soria N, et al. *RUNX1/RUNX1T1* controls alternative splicing in the t(8;21)-positive acute myeloid leukemia cells. *BioRxiv*. 2019. p. 628040. DOI: 10.1101/628040.
22. Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcriptase template switching and false alternative transcripts. *Genomics*. 2006;88(1):127–131. DOI: 10.1016/j.ygeno.2005.12.013.
23. Houseley J, Tollervey D. Apparent Non-Canonical Trans-Splicing is generated by reverse transcriptase *in vitro*. *PLOS ONE*. 2010;5(8):e12271. DOI: 10.1371/journal.pone.0012271.
24. Markova EN, Kantidze OL, Razin SV. Transcription of the AML1/ETO chimera is guided by the P2 promoter of the AML1 gene in the Kasumi-1 cell line. *Gene*. 2012;510(2):142–146. DOI: 10.1016/j.gene.2012.09.028.
25. Kozu T, Fukuyama T, Yamami T, Akagi K, Kaneko Ya. MYND-less splice variants of AML1-MTG8 (*RUNX1-CBFA2T1*) are expressed in leukemia with t(8;21). *Genes Chromosomes Cancer*. 2005;43(1):45–53. DOI: 10.1002/gcc.20165.
26. Wan Y, Larson DR. Splicing heterogeneity: separating signal from noise. *Genome Biology*. 2018;19(1):86. DOI: 10.1186/s13059-018-1467-4.
27. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' UTRs and fewer microRNA target sites. *Science*. 2008;320(5883):1643–1647. DOI: 10.1126/science.1155390.
28. Johnson DT, Shima T, Davis AG, Zhang D-E. Characterization of the Post-Transcriptional Regulation of AML1-ETO Expression in t(8;21) Leukemia Cells. *Blood*. 2017;130(1):3790. URL: [http://www.bloodjournal.org/content/130/Suppl\\_1/3790](http://www.bloodjournal.org/content/130/Suppl_1/3790).
29. Junge A, Zandi R, Havgaard JH, Gorodkin J, Cowland JB. Assessing the miRNA sponge potential of *RUNX1T1* in t(8;21) acute myeloid leukemia. *Gene*. 2017;615(C):35–40. DOI: 10.1016/j.gene.2017.03.015.

## References

- Peterson LF, Boyapati A, Ahn E, Biggs JR, Okumura AJ, Lo M-C, et al. Acute myeloid leukemia with the 8q22;21q22 translocation: secondary mutational events and alternative t(8;21) transcripts. *Blood*. 2007;110(3):799–805. DOI: 10.1182/blood-2006-11-019265.
- LaFiura KM, Edwards H, Taub JW, Matherly LH, Fontana JA, Mohamed AN, et al. Identification and characterization of novel AML1-ETO fusion transcripts in pediatric t(8;21) acute myeloid leukemia: a report from the Children's Oncology Group. *Oncogene*. 2008;27(36):4933–4942. DOI: 10.1038/onc.2008.134.
- Grinev VV, Migas AA, Kirsanova AD, Mishkova OA, Siomava N, Ramanouskaya TV, et al. Decoding of exon splicing patterns in the human *RUNX1-RUNX1T1* fusion gene. *International Journal of Biochemistry and Cell Biology*. 2015;68:48–58. DOI: 10.1016/j.biocel.2015.08.017.
- Migas AA, Mishkova OA, Ramanouskaya TV, Ilyushonak IM, Aleinikova OV, Grinev VV. *RUNX1T1/MTG8/ETO* gene expression status in human t(8;21)(q22;q22)-positive acute myeloid leukemia cells. *Leukemia Research*. 2014;38(9):1102–1110. DOI: 10.1016/j.leukres.2014.06.002.
- Pfaffl MW. A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Research*. 2001;29(9):e45. DOI: 10.1093/nar/29.9.e45.
- Ramakers C, Ruijter JM, Deprez RHL, Moorman AF. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neuroscience Letters*. 2003;339(1):62–66. DOI: 10.1016/S0304-3940(02)01423-4.
- Ruijter JM, Ramakers C, Hoogaars WM, Karlen Y, Bakker O, van den Hoff MJ, et al. Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Research*. 2009;37(6):e45. DOI: 10.1093/nar/gkp045.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–2120. DOI: 10.1093/bioinformatics/btu170.
- Birol I, Raymond A, Chiu R, Nip KM, Jackman SD, Kreitzman M, et al. Kleat: cleavage site analysis of transcriptomes. In: Altman RB, Dunker AK, Hunter L, Ritchie MD, Murray TA, Klein TE, editors. *Biocomputing-2015. Pacific Symposium on Biocomputing; 2015 January 4–8; Fairmont Orchid, Hawaii*. [S. l.]: [s. n.]; 2015. p. 347–358. DOI: 10.1142/9789814644730\_0034.
- Xiao Z, Greaves MF, Buffler P, Smith M. Molecular characterization of genomic AML1-ETO fusions in childhood leukemia. *Leukemia*. 2001;15(12):1906–1913. DOI: 10.1038/sj.leu.2402318.
- Zhang Y, Strissel P, Strick R, Chen J, Nucifora G, Le Beau MM, et al. Genomic DNA breakpoints in AML1/RUNX1 and ETO cluster with topoisomerase II DNA cleavage and DNase I hypersensitive sites in t(8;21) leukemia. *Proceedings of the National Academy of Sciences*. 2002;99(5):3070–3075. DOI: 10.1073/pnas.042702899.
- Kent WJ. BLAT – the BLAST-like alignment tool. *Genome Research*. 2002;12(4):656–664. DOI: 10.1101/gr.229202.
- Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, et al. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res*. 2016;44(D1):D717–725. DOI: 10.1093/nar/gkv1275.
- Ilyushonak IM, Gunko EP, Antonovich ML, Yatskou MM, Kustanovich AM, Sukhareuski AY, et al. Study of RNA splicing patterns of the human *RUNX1-RUNX1T1* fusion oncogene by the methods of data mining and high-throughput DNA sequencing. In: Kil'chevskii AV, chief editor. *Molekulyarnaya i prikladnaya genetika. Sbornik nauchnykh trudov. Tom 23*. Minsk: Institut genetiki i tsitologii NAN Belarusi; 2017. p. 92–101. Russian.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nature Biotechnology*. 2011;29(1):24–26. DOI: 10.1038/nbt.1754.
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*. 2017;14(4):417–419. DOI: 10.1038/nmeth.4197.
- Pages H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.30.1. 2014. DOI: 10.18129/B9.bioc.Biostrings.
- Morgan M, Pages H, Obenchain V, Hayden N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 1.34.1. 2019. DOI: 10.18129/B9.bioc.Rsamtools.
- Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*. 2013;9(8):e1003118. DOI: 10.1371/journal.pcbi.1003118.
- Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*. 2011;6:26. DOI: 10.1186/1748-7188-6-26.
- Grinev VV, Ilyushonak IM, Clough R, Nakjang S, Smink J, Martinez-Soria N, et al. *RUNX1/RUNX1T1* controls alternative splicing in the t(8;21)-positive acute myeloid leukemia cells. *BioRxiv*. 2019. p. 628040. DOI: 10.1101/628040.
- Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcriptase template switching and false alternative transcripts. *Genomics*. 2006;88(1):127–131. DOI: 10.1016/j.ygeno.2005.12.013.
- Houseley J, Tollervey D. Apparent Non-Canonical Trans-Splicing is generated by reverse transcriptase *in vitro*. *PLOS ONE*. 2010;5(8):e12271. DOI: 10.1371/journal.pone.0012271.
- Markova EN, Kantidze OL, Razin SV. Transcription of the AML1/ETO chimera is guided by the P2 promoter of the AML1 gene in the Kasumi-1 cell line. *Gene*. 2012;510(2):142–146. DOI: 10.1016/j.gene.2012.09.028.
- Kozu T, Fukuyama T, Yamami T, Akagi K, Kaneko Ya. MYND-less splice variants of AML1-MTG8 (*RUNX1-CBFA2T1*) are expressed in leukemia with t(8;21). *Genes Chromosomes Cancer*. 2005;43(1):45–53. DOI: 10.1002/gcc.20165.
- Wan Y, Larson DR. Splicing heterogeneity: separating signal from noise. *Genome Biology*. 2018;19(1):86. DOI: 10.1186/s13059-018-1467-4.
- Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' UTRs and fewer microRNA target sites. *Science*. 2008;320(5883):1643–1647. DOI: 10.1126/science.1155390.
- Johnson DT, Shima T, Davis AG, Zhang D-E. Characterization of the Post-Transcriptional Regulation of AML1-ETO Expression in t(8;21) Leukemia Cells. *Blood*. 2017;130(1):3790. URL: [http://www.bloodjournal.org/content/130/Suppl\\_1/3790](http://www.bloodjournal.org/content/130/Suppl_1/3790).
- Junge A, Zandi R, Havgaard JH, Gorodkin J, Cowland JB. Assessing the miRNA sponge potential of *RUNX1T1* in t(8;21) acute myeloid leukemia. *Gene*. 2017;615(C):35–40. DOI: 10.1016/j.gene.2017.03.015.