На правах рукописи

ЛЕТЯГИНА АННА ЕВГЕНЬЕВНА

АНАЛИЗ ВЛИЯНИЯ ПОСЛЕДОВАТЕЛЬНОСТИ, РАСПОЛОЖЕННОЙ ПОСЛЕ САЙТА ПОЛИАДЕНИЛИРОВАНИЯ, НА УРОВЕНЬ ЗРЕЛОЙ МРНК РЕПОРТЁРНОГО ГЕНА eGFP В КУЛЬТИВИРУЕМЫХ КЛЕТКАХ ЧЕЛОВЕКА HEK293T

1.5.7 – Генетика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата биологических наук

Новосибирск – 2025

Работа выполнена в лаборатории клеточного деления Института молекулярной и клеточной биологии СО РАН и на кафедре цитологии и генетики факультета естественных наук Федерального государственного автономного образовательного учреждения высшего образования «Новосибирский национальный исследовательский государственный университет», г. Новосибирск.

Научный руководитель:	**Омелина Евгения Сергеевна** кандидат биологических наук, заведующая лабораторией клеточного деления Института молекулярной и клеточной биологии СО РАН, г. Новосибирск
Официальные оппоненты:	**Максименко Оксана Геннадьевна** доктор биологических наук, руководитель центра высокоточного редактирования и генетических технологий для биомедицины «ФГБУН Институт биологии гена РАН», г. Москва
	Скоблов Михаил Юрьевич Кандидат биологических наук, доцент, заведующий отделом функциональной геномики ФГБНУ «Медико-генетический научный центр имени академика Н.П. Бочкова РАМН», г. Москва
Ведущее учреждение:	ФГАОУ ВО Первый МГМУ имени И.М. Сеченова Минздрава России (Сеченовский университет), г. Москва

Защита диссертации состоится «11» февраля 2026 г. на дневном заседании Диссертационного совета 24.1.239.01 на базе ФГБНУ «Федерадьный исследовательсктй центр Институт цитологии и генетики Сибирского отделения РАН» в конференц-зале Института по адресу: 630090, г. Новосибирск, проспект ак. Лаврентьева, 10, тел. (383) 363-49-06, факс (383) 333-12-78, e-mail: dissov@bionet.nsc.ru.

С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте Института http://www.icgbio.ru.

Автореферат разослан «__» _____ 2025 г.

Учёный секретарь
диссертационного совета,
доктор биологических наук Хлебодарова Т.М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность исследования. Экспрессия генов является определяющим процессом в жизнедеятельности всех организмов, который сложным образом регулируется на всех его этапах. У эукариотических организмов основной вклад в активность генов вносят структура хроматина, определяемая взаимодействиями регуляторных белков со специфическими мотивами ДНК, а также процессинг (созревание) РНК. Исследование влияния различных последовательностей ДНК на регуляцию экспрессии гена является актуальной задачей генетики и биотехнологии. При этом роль регуляторных районов, расположенных выше промоторных последовательностей, в этом процессе изучена значительно лучше, чем районов, расположенных в 3'-областях генов.

У эукариот терминация транскрипции может оказывать значительное влияние на уровень экспрессии генов различными путями. Процессинг 3'-конца предшественника матричной РНК (пре-мРНК) и последующая терминация транскрипции белок-кодирующих генов основана на сборке функционального комплекса процессинга 3'-конца пре-мРНК. Нуклеотидный состав 3'-участка незрелого транскрипта влияет на то, какой именно вариант комплекса процессинга 3'-конца пре-мРНК будет сформирован и какой сигнал полиаденилирования (СПА) будет выбран. Нуклеотидные последовательности, индуцирующие и улучшающие сборку комплекса процессинга 3'-конца пре-мРНК, исследуются на протяжении последних 49 лет (Proudfoot and Brownlee, 1976). Однако информация о составе оптимальной последовательности Downstream sequence element (DSE) – одного из ключевых цис-элементов, принимающих участие в процессинге 3'-конца пре-мРНК, остаётся неполной и противоречивой. Таким образом, детальное исследование механизмов терминации транскрипции с учётом регуляторной активности нуклеотидных мотивов на 3'-конце гена актуально для понимания регуляции экспрессии белок-кодирующих генов у млекопитающих.

Данная работа направлена на исследование влияния последовательности, расположенной после СПА (DSE), на уровень зрелой мРНК в культивируемых клетках почки эмбриона человека HEK293T. В рамках работы был исследован уровень экспрессии репортёрных плазмидных конструкций, несущих ген улучшенного зелёного флуоресцентного белка (eGFP) под контролем промотора фосфоглицераткиназы 1 человека (hPGK) и СПА растворимого нейропилина 1 (sNRP-1), за которым располагался 3'-терминальный повтор (TR) транспозона PiggyBac. DSE в данной конструкции расположен в области PiggyBac 3'-TR.

Степень разработанности проблемы. За последние 15 лет было опубликовано несколько работ, показывающих, что терминаторы транскрипции РНК-полимеразы II не только обеспечивают процессинг 3'-

конца предшественника матричной РНК (пре-мРНК) и диссоциацию тройного комплекса, но и регулируют уровень зрелой мРНК (West and Proudfoot, 2009; Bogard *et al*., 2019; Vainberg Slutskin *et al*., 2019; Wang *et al*., 2022; Zhou *et al*., 2023).

Ранее сотрудниками нашей лаборатории было показано, что делеция одного цитозина в позиции +32 п.н. ниже СПА sNRP-1 (ΔC) приводит к двукратному повышению уровня транскриптов репортёрного гена eGFP в культивируемых эмбриональных стволовых клетках мыши mESCs, мышиных клетках 3T3 и клетках человека HEK293T (Akthar *et al*., 2013; Boldyreva *et al*., 2021). Таким образом, можно предположить, что последовательность ΔC вовлечена в некий консервативный для млекопитающих молекулярный механизм. Было показано, что изменение последовательности в области +17..56 п.н. после СПА sNRP-1 влияет на уровень зрелой мРНК и белка eGFP (Boldyreva *et al*., 2021). Область +17..56 п.н. после СПА в исследованной конструкции по своему расположению соответствует области элемента последовательности, расположенного ниже СПА (DSE).

Цели и задачи работы. Цель: исследовать влияние последовательности, расположенной после сайта полиаденилирования, на уровень зрелой мРНК гена eGFP в культивируемых клетках почки эмбриона человека HEK293T.

Задачи:

1. Определить район после сайта полиаденилирования, оказывающий наибольшее влияние на уровень зрелой мРНК репортёрного гена eGFP;

2. Выделить признаки интересующей последовательности, оказывающие наибольшее влияние на уровень зрелой мРНК гена eGFP;

3. Исследовать влияние стабильности предсказанной вторичной структуры пре-мРНК в области после СПА на уровень зрелой мРНК eGFP;

4. Разработать новые терминаторы транскрипции, позволяющие модулировать уровень зрелой мРНК и белка eGFP.

Научная новизна. В рамках работы будет впервые системно исследовано влияние последовательности DSE и стабильности предсказанной вторичной структуры пре-мРНК после сигнала полиаденилирования на уровень зрелой мРНК eGFP в культивируемых клетках HEK293T.

Теоретическая и практическая значимость работы. Результаты данной работы имеют важное фундаментальное значение в понимании регуляции экспрессии генов эукариот, а также практическое применение в плане развития генетической инженерии и биотехнологии.

Методология и методы исследования. Одной из причин малой исследованности регуляторной роли терминаторов транскрипции является отсутствие технических подходов, позволяющих систематически

идентифицировать функциональные элементы, расположенные СПА и не входящие в последовательности зрелых (полиаденилированных) транскриптов. В данной работе использован метод массового параллельного репортёрного анализа (МПРА), который позволяет одновременно измерять уровень транскрипционной активности большого числа (до нескольких миллионов) не интегрированных в геном трансгенов. Метод МПРА основан на кратковременной трансфекции культивируемых клеток ДНК-штрихкодированными репортёрными конструкциями с последующим анализом их транскрипционной активности посредством высокопроизводительного параллельного секвенирования. Метод МПРА был использован для систематического диссекционного анализа DSE. DSE расположен на расстоянии 10-30 н. ниже сайта полиаденилирования и обеспечивает связывание белка CstF64 с молекулой пре-мРНК. DSE не входит в состав молекулы зрелой мРНК. Это делает его привлекательной мишенью для модуляции уровня экспрессии целевого гена.

Положения, выносимые на защиту:

1. Район, расположенный в терминальной области гена eGFP на расстоянии +17..+40 п.н. от сигнала полиаденилирования наиболее сильно влияет на уровень его зрелой мРНК в культивируемых клетках человека HEK293T.

2. Уровень зрелой мРНК гена eGFP в клетках HEK293T коррелирует с уровнем предсказанной минимальной свободной энергии вторичной структуры, возникающей в последовательности пре-мРНК после сигнала полиаденилирования.

Апробация результатов. Результаты работы представлены и обсуждены на международной конференции «Хромосома» в 2018 и 2023 гг. (Новосибирск), на международной конференции «BGRS\SB-2018» (Новосибирск), на международной мини-конференции «Chromosomes and Mitosis» в 2019 г. (Новосибирск), на VII (2019 г., Санкт-Петербург) и VIII (2024 г., Саратов) Съездах ВОГиС, на II Объединённом научном форуме в 2019 г. (Москва), на IX Всероссийском молодёжном научном форуме в 2024 г. (Самара). По теме диссертации были опубликованы 3 работы в рецензируемых научных журналах, входящих в международные базы цитирования (WoS, Scopus, РИНЦ).

Структура и объём диссертации. Диссертация включает в себя введение, обзор литературы, материалы и методы, результаты, заключение, выводы, благодарности, список сокращений и условных обозначений и список литературы (535 источников). Общий объём составляет 208 страниц, в том числе 10 таблиц, 29 рисунков и 2 приложения.

МАТЕРИАЛЫ И МЕТОДЫ

Получение плазмидных библиотек для МПРА. Плазмида pTTC-hPGK-eGFP была получена из плазмиды pTTC-Hsap-WT (Boldyreva *et al*., 2021) вырезанием прследовательности гена mCherry методом рестрикции-лигирования. Библиотеки для МПРА конструировали на основе плазмиды pTTC-hPGK-eGFP с помощью метода сборки Гибсона. В вектор были встроены случайные последовательности: 18-тибуквенный штрихкод и 8-мибуквенная мутация в положении -59..-42 п.н. выше и +17..+56 п.н. ниже СПА sNRP-1 соответственно.

Ведение клеточных культур. Культивируемые клетки почки эмбриона человека (HEK293T) были получены из банка данных ATCC. Культивируемые клетки HEK293T вели при 37°C в увлажнённом воздухе с 5% содержанием CO_2 в среде Игла, модифицированной Дульбекком (DMEM, Gibco) с добавлением фетальной бычьей сыворотки (FBS, Gibco) до 10%, 7.5% NaHCO3, 100 МЕ/мл пенициллина и 100 мкг/мл стрептомицина.

Временная трансфекция клеток. За день до трансфекции высевали по $0.5\text{-}1\times10^6$ культивируемых клеток в лунку шестилуночного планшета (9.6 $см^2$). Трансфицировали культивируемые клетки 400 нг плазмид с использованием реагента для трансфекции Effectene (Qiagen) или GenJect-40 (Молекта) согласно рекомендациям производителей. После 48 часов инкубации собирали клетки для выделения тотальной РНК или проведения FACS анализа.

Выделение тотальной РНК и синтез кДНК. Трансфицированные культивируемые клетки HEK293T лизировали в 1 мл RNAzol RT (Molecular Research Center) и выделяли тотальную РНК согласно рекомендациям производителя. Очищенную РНК инкубировали с ДНКазой I (Thermo Fisher Scientific) и эндонуклеазой рестрикции DpnI (New England Biolabs) чтобы избавиться от примесей гДНК и плазмидной ДНК (пДНК). Затем очищали РНК с помощью набора CleanRNA Standard (Евроген) согласно рекомендациям производителя. Реакцию обратной транскрипции проводили с использованием 1-3 мкг очищенной тотальной РНК в качестве матрицы.

Получение образцов картирования, нормирования и экспрессии. Подготовка всех образцов МПРА включала в себя 2 раунда ПЦР. Во время I раунда каждый образец МПРА метили уникальным восьмибуквенным индексом. Во время II раунда ПЦР к образцам присоединяли адаптеры P5 и P7 для секвенирования на платформе Illumina MiSeq. Для получения образцов картирования в I раунде в качестве матрицы использовали 10 пг пДНК, во II раунде – 0.5 мкл очищенных продуктов первого раунда ПЦР. Для получения образцов нормирования в I раунде в качестве матрицы использовали 2.5 нг пДНК, во II раунде – 2 мкл реакционной смеси I раунда ПЦР. Для получения образцов экспрессии в I раунде в качестве матрицы использовали 4 мкл кДНК

(1/5 объёма реакционной смеси для проведения ОТ), во II раунде – 4 мкл реакционной смеси I раунда ПЦР.

Проведение количественной ПЦР в реальном времени (кПЦР). кПЦР проводили с использованием набора БиоМастер HS-qPCR SYBR Blue (Биолабмикс) и амплификатора CFX96 Touch Real-Time PCR Detection System (Bio-Rad). Детекцию проводили на каждом раунде амплификации.

Секвенирование образцов МПРА и анализ данных. Для каждой плазмидной библиотеки были получены минимум 2 образца картирования, нормирования и экспрессии, меченные различными индексами. Все образцы МПРА для двух плазмидных библиотек смешивали между собой и секвенировали на платформе Illumina MiSeq. Образцы секвенировали в одном направлении, длина прочтения составляла 151 н. Длина прочтения была меньше длины секвенируемых продуктов ПЦР, поэтому не было необходимости в удалении последовательностей адаптеров из полученных в результате секвенирования последовательностей. Данные секвенирования образцов МПРА были проанализированы с помощью собственной программы MPRAdecoder (Letiagina *et al*., 2021), написанной на языке Python.

Сортировка клеток, активированная флуоресценцией (FACS). Трансфицированные клетки собирали для анализа FACS через 48-72 часа после трансфекции. Среднюю интенсивность флуоресценции измеряли при 510 нм (eGFP) и 640 нм (mCherry) с использованием проточного цитометра NovoCyte (Agilent).

Предсказание минимальной свободной энергии (МСЭ) вторичной структуры РНК. Для расчёта МСЭ района пре-мРНК, содержащего DSE, был выбран фрагмент пре-мРНК, включающий в себя 37 н. до основного сайта полиаденилирования и 25 н. после него. Для расчёта МСЭ района пре-мРНК, включающего в себя штрихкод, был выбран фрагмент пре-мРНК, расположенный в районе -92..-32 н. выше СПА. МСЭ была предсказана с помощью программы RNAfold пакета ViennaRNA (Lorenz *et al*., 2011).

Обучение моделей, предсказывающих уровень зрелой мРНК на основе последовательности, расположенной после СПА и последовательности штрихкода. Все модели были получены с использованием методов классического машинного обучения. Программа для обучения моделей была написана на языке Python и использовала библиотеку Pycaret (https://pycaret.readthedocs.io/en/latest/). Обученные модели доступны по ссылке https://github.com/AnnLetiagina/DoSIA.

Дизайн новых последовательностей. Для предсказания новых последовательностей была использована собственная программа DSEgenerator. Она доступна по ссылке https://github.com/AnnLetiagina/DoSIA/tree/main/Example.

РЕЗУЛЬТАТЫ

Дизайн библиотек для МПРА. Метод МПРА основан на использовании библиотек, содержащих два ключевых фрагмента: исследуемую последовательность (в данной работе условно названную «мутация») и штрихкод (ШК). ШК и исследуемая последовательность обычно представляют собой короткие последовательности, расположенные внутри и вне транскрипционной единицы, соответственно. Таким образом, ШК могут быть использованы для количественной оценки влияния различных изучаемых последовательностей, отсутствующих в зрелых транскриптах, на представленность последних в трансфицированных клетках (Tewhey *et al.*, 2016; Komura *et al.*, 2018).

Данное исследование направлено на систематический анализ влияния нуклеотидного состава последовательности, расположенной после СПА, на уровень зрелой мРНК вышележащего репортёрного гена eGFP в клетках человека HEK293T. Для этого н.с. лаборатории клеточного деления ИМКБ СО РАН Л.А. Яринич были сконструированы девять плазмидных библиотек для МПРА, в которых ШК и мутация располагались в 3’-НТО репортёрного гена eGFP и после СПА sNRP-1, соответственно (Рисунок 1). Мутации представляли собой взаимно перекрывающиеся последовательности длиной 8 п.н., расположенные в позициях +17..56 п.н. после СПА. Мутации вводили в плазмидный вектор с помощью случайных олигонуклеотидных праймеров. ШК также представляли собой случайные последовательности, но значительно большей длины (18 п.н.). Плазмидные библиотеки были названы в соответствии с положением мутации. Например, в плазмидах библиотеки 17-24 мутация (исследуемый вариабельный участок) расположена в позиции +17..24 п.н. после СПА. Для нормирования в каждую плазмидную библиотеку для МПРА добавляли в соотношении 1/100 эквимолярный пул из двух контрольных конструкций с исходной последовательностью терминатора транскрипции "дикого типа" (WT) и двух контрольных конструкций с делецией цитозина в позиции +32 н. ниже СПА (ΔC), меченных специально разработанными 20-буквенными ШК. Клетки HEK293T, трансфицированные библиотеками для МПРА, собирали для оценки количества ШК в транскриптах eGFP через 48 ч после трансфекции.

Оптимизация условий ПЦР позволяет снизить долю химерных продуктов до 0.3%. Поскольку мы использовали МПРА-библиотеки с заранее неизвестными последовательностями ШК и мутаций, было необходимо подготовить т.н. образцы картирования, чтобы определить уникальные сочетания ШК-мутация (Рисунок 2). Для выявления всех уникальных комбинаций ШК-мутация обычно используется ПЦР-амплификация с последующим высокопроизводительным секвенированием. Однако ранее было показано, что традиционная ПЦР-коамплификация последовательностей ДНК, содержащих два вариабельных мотива (ШК и мутацию в случае МПРА),

разделенных константной областью, часто приводит к образованию нежелательных химерных молекул, в количестве от 5.4 до 30% (Acinas *et al.*, 2005: Shao *et al.*, 2011; Bjørnsgaard *et al.*, 2017; Porapov and Ong, 2017).

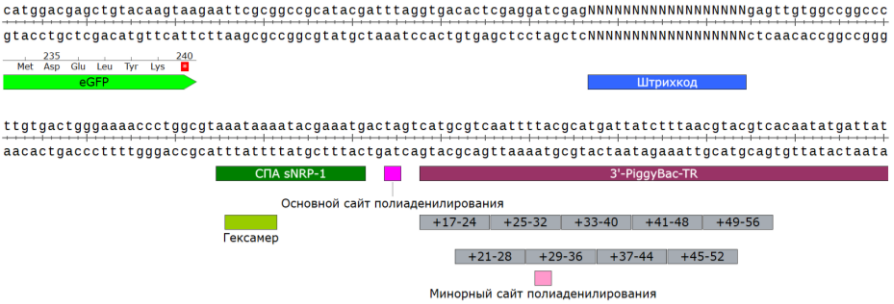

Рисунок 1. Структура библиотек для МПРА (на примере последовательности конструкции WT), использованных для оценки влияния на экспрессию репортёрного гена eGFP последовательностей, расположенных после СПА. Мутации изображены как серые прямоугольники.

Такие продукты ПЦР усложняют и могут вводить в заблуждение при анализе данных МПРА, а также снижают производительность подхода, поскольку ассоциация одного и того же ШК с разными исследуемыми последовательностями приводит к исключению всех таких ШК и исследуемых последовательностей из анализа. Химерные молекулы ДНК, по-видимому, образуются в результате отжига неполностью удлиненных праймеров к гетерологичной целевой последовательности во время ПЦР (Bradley and Hillis, 1997). Неполное удлинение нитей ДНК, предположительно, является следствием остановки ДНК-полимеразы на матрице или ее преждевременной терминации. Репликация гетерологичного дуплекса приводит к образованию химерных продуктов ПЦР, состоящих из двух искусственно комбинированных последовательностей (Klug *et al.*, 1991).

Частота образования химерных молекул зависит от длины и сходства последовательностей совместно амплифицируемых молекул ДНК (Wang and Wang, 1996). Кроме того, известно, что количество ДНК-матрицы, число циклов амплификации, размер плазмидной библиотеки и продолжительность этапа элонгации играют решающую роль в образовании химерных ПЦР-продуктов (Lahr and Katz, 2009; Liu *et al.*, 2014). Однако считается, что основной причиной образования химерных молекул является присутствие разных матриц ДНК в одной реакционной смеси ПЦР (Shao *et al.*, 2011; Boers *et al.*, 2015). Таким образом, использование метода эмульсионной ПЦР (эПЦР), обеспечивающего простое физическое разделение молекул ДНК-матрицы за счёт использования эмульсии "вода в масле" (Williams *et al.*, 2006), представляется прекрасным решением этой проблемы.

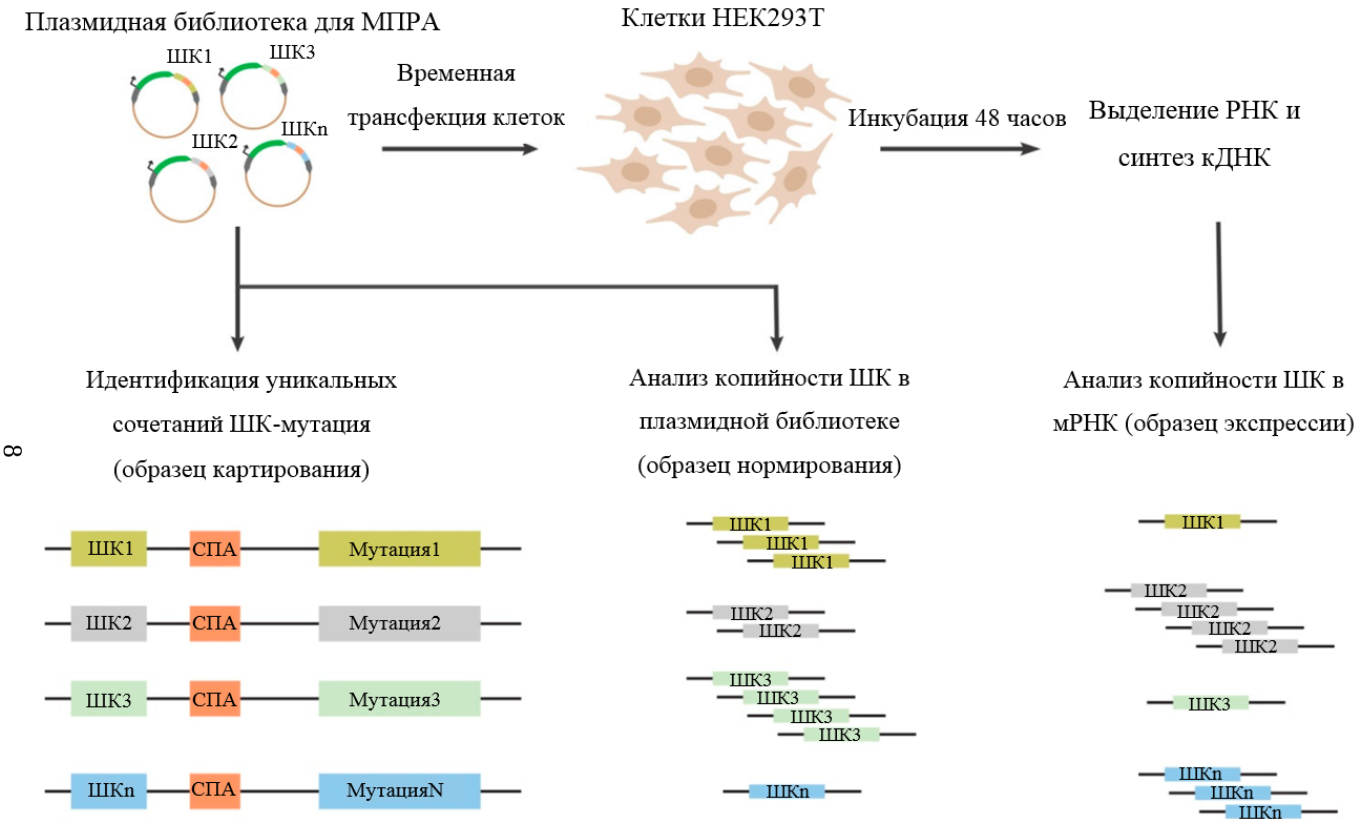

Рисунок 2. Схема экспериментальных этапов приготовления образцов МПРА для последующего высокопроизводительного секвенирования в соответствии с Omelina *et al*., 2019. ШК – штрихкод; СПА – сигнал полиаденилирования.

Поэтому для наработки образцов картирования мы использовали набор Micellula DNA Emulsion & Purification Kit (EURX). В ходе многоэтапной оптимизации были подобраны оптимальные условия ПЦР, позволяющие снизить долю химерных продуктов до 0.3%. Долю химерных продуктов вычисляли с помощью анализа данных высокопроизводительного секвенирования продуктов II раунда ПЦР с помощью собственной программы MPRAdecoder (Letiagina *et al.*, 2021). Для получения образцов картирования в I раунде в качестве матрицы использовали 10 пг ДНК одной из плазмидных библиотек, во II раунде – 0.5 мкл очищенных продуктов первого раунда ПЦР. В I раунде ПЦР было 15 циклов амплификации, во II – 18 циклов амплификации. Время элонгации составляло 30 с в обоих раундах ПЦР.

Далее, чтобы напрямую сравнить эффективность эмульсионной и обычной ПЦР, мы повторили амплификацию фрагментов ШК-мутация плазмидных библиотек с помощью обычной ПЦР с теми же настройками, которые использовались для эПЦР. Неожиданно оказалось, что доля химерных молекул ШК-мутация в продуктах обычной ПЦР (0.32 %) лишь немного выше, чем в продуктах эПЦР. Таким образом, мы пришли к выводу, что и эмульсионная, и традиционная ПЦР с оптимизированными настройками могут быть успешно использованы для эффективной идентификации изначально неизвестных сочетаний ШК-мутация, присутствующих в плазмидных библиотеках для МПРА.

Подготовка образцов для МПРА. Так как обычная ПЦР при оптимальных условиях амплификации позволяет снижать долю химерных продуктов почти так же эффективно, как и эПЦР, для подготовки образцов картирования для МПРА была проведена двухраундовая традиционная ПЦР для амплификации фрагментов ШК-мутация с использованием плазмидных библиотек в качестве матрицы (Рисунок 2).

Образцы экспрессии и нормирования были подготовлены путем ПЦР-амплификации последовательностей ШК с использованием кДНК, полученной из трансфицированных клеток HEK293T, и ДНК плазмидной библиотеки в качестве матрицы, соответственно (Рисунок 2). Соотношение представленности каждого ШК в образцах экспрессии и нормирования позволяет судить о влиянии соответствующего варианта последовательности мутации на экспрессию репортёрного гена eGFP. Образцы картирования, нормирования и экспрессии были проанализированы с помощью секвенирования на платформе Illumina MiSeq. Для каждого повтора образца в среднем было получено ~1.5 млн одноконцевых прочтений длиной 151 н. Для каждой плазмидной библиотеки было проанализировано минимум 2 повтора нормирования, картирования и экспрессии. Для каждого ШК вычислялись средние значения представленности, измеренные в разных повторах нормирования и картирования.

Анализ данных МПРА. ***Наибольшее влияние на уровень зрелой мРНК оказывают последовательности, расположенные в районе +17..+40 н. после СПА.*** Анализ данных высокопроизводительного секвенирования был проведён с помощью собственной программы MPRAdecoder (Letiagina *et al.*, 2021). Было обнаружено, что более 40% мутаций в районах +17...24, +25...32, +29...36 и +33...40 после СПА приводят более чем к двукратному увеличению уровня зрелой мРНК eGFP (Таблица 1). Кроме того, для этих районов было характерно наличие мутаций, увеличивающих уровень зрелой мРНК eGFP более чем в 5 раз. В районах +41...48, +45...52 и +49...56 не более 1.4% мутаций повышают или понижают уровень зрелой мРНК eGFP более чем в 2 раза по сравнению с исходной последовательностью. Таким образом, последовательности, расположенные ниже +41 п.н. после СПА, оказывают минимальное влияние на уровень зрелой мРНК eGFP.

Таблица 1. Влияние мутаций и ШК на уровень зрелой мРНК eGFP в плазмидных библиотеках для МПРА. В таблице указано логарифмическое значение нормированного уровня зрелой мРНК eGFP. Столбец «=>1» отражает долю последовательностей мутаций, повышающих уровень зрелой мРНК eGFP более чем в 2 раза по сравнению с исходной последовательностью. Столбец «<=-1» отражает долю последовательностей мутаций, понижающих уровень зрелой мРНК eGFP более чем в 2 раза по сравнению с исходной последовательностью. В нижней строке приведены данные о влиянии последовательностей ШК. Макс. – максимальная; Мин. – минимальная; Ср. – средняя; ШК – штрихкоды; Экспр. – экспрессия (уровень зрелой мРНК eGFP); ρ – коэффициент корреляции Пирсона.

Библиотека	Макс. экспр., log_2	Мин. экспр., log_2	Ср. экспр., log_2	=>1, %, log_2	<=-1, %, log_2	ρ для мутаций с двумя и более ШК
17-24	2.55	-1.6	0.9	43	0.1	0.43
21-28	1.96	-1.89	0.57	15	0.1	0.63
25-32	3.76	-3.13	0.94	48	2	0.47
29-36	3.53	-1.93	1.35	70	0.05	0.71
33-40	4.26	-3.02	1.59	78	0.3	0.52
37-44	2.38	-2.4	0.45	11	1	0.37
41-48	2.17	-2.02	0.07	0.4	1	0.19
45-52	1.36	-2.31	0.09	0.1	0.1	0.27
49-56	1.81	-2.54	0.004	0.2	0.6	0.06
49-56 + WT (ШК)	2.39	-2.86	0.06	0.02	0.04	-

Известно, что последовательности, расположенные выше СПА, могут оказывать влияние на сборку канонического комплекса процессинга 3'-конца пре-мРНК (Lesnik and Frier, 1995; Hu *et al.*, 2005). Таким образом, последовательности ШК, наряду с мутациями, могут влиять на уровень зрелой мРНК eGFP. Для мутаций, ассоциированных с двумя и более ШК, влияние последних нивелируется усреднением значений уровней зрелой мРНК eGFP, ассоциированных с разными ШК. Анализ данных высокопроизводительного секвенирования показал, что в каждой нашей библиотеке для МПРА присутствуют плазмиды, имеющие ШК, но на месте мутации сохранившие исходную последовательность после СПА (WT). Для того, чтобы удостовериться в том, что последовательности мутаций действительно влияют на уровень зрелой мРНК eGFP, мы проанализировали выборку таких ШК, ассоциированных с исходной последовательностью, и сравнили её с общим пулом ШК. Доля таких ШК составляет в среднем 1.2% от всех ШК плазмидной библиотеки. Вариабельность уровней зрелой мРНК в этой выборке обусловлена только влиянием ШК, т.к. последовательности мутаций в них идентичны. Для того, чтобы проверить, отличается ли выборка уровней распределения зрелой мРНК eGFP для ШК, ассоциированных с исходной последовательностью, от выборки уровней распределений зрелой мРНК eGFP для ШК, ассоциированных с разнообразными мутациями, мы использовали U-критерий Манна-Уитни. Анализ показал, что только последовательность, расположенная в районе +49..56 п.н. после СПА, не оказывает влияния на уровень зрелой мРНК eGFP. Это подтверждается низким коэффициентом корреляции Пирсона – 0.06 – для мутаций с двумя и более штрихкодами в библиотеке 49-56 (Таблица 1).

Оказалось, что для 5% мутаций, обеспечивающих наиболее высокий уровень экспрессии eGFP, характерно выраженное обогащение тимином. С помощью точного критерия Фишера было показано, что мутации с уровнем экспрессии eGFP, превышающим уровень экспрессии исходной последовательности, были обогащены мотивами CTC, TCT, TGT, TTG, GTCT, GTGT, TGTC, TCTC, характерными для сайтов связывания белка CstF64 (Beyer *et al*., 1997; Pérez-Cañadillas and Varani, 2003; Monarez *et al*., 2007), тогда как мутации с уровнем экспрессии eGFP ниже, чем у исходной последовательности, были обогащены мотивами AGG, GGG, GGC, AGGG, CAGG, GGCC, AGGC.

Низкая стабильность вторичной структуры после СПА коррелирует с высоким уровнем экспрессии репортёра. Также был проведён анализ влияния стабильности вторичной структуры пре-мРНК в районе после СПА на уровень экспрессии eGFP. Была проанализирована вторичная структура фрагмента пре-мРНК, включающего в себя 37 н. до основного сайта полиаденилирования и 25 н. после него. У непроцессированных транскриптов гена eGFP этот фрагмент пре-мРНК включал в себя СПА, сайт

полиаденилирования и район до 56 н. после СПА. Оказалось, что непроцессированные транскрипты исходной последовательности имеют вторичную структуру после СПА. Для того, чтобы проверить, влияет ли стабильность предсказанной вторичной структуры пре-мРНК в районе после СПА на уровень зрелой мРНК eGFP, была предсказана МСЭ всех участков пре-мРНК после СПА, включающих в себя исследованные в ходе МПРА мутации в соответствующих позициях после СПА. Расчёт коэффициента корреляции Пирсона показал, что для библиотек 17-24, 21-28, 25-32, 29-36, 33 40, 37-44 и 41-48 повышение МСЭ достоверно коррелирует с повышением уровня экспрессии eGFP (Таблица 2).

Таблица 2. Понижение стабильности вторичной структуры пре-мРНК в районе +21..+48 н. после СПА статистически достоверно связано с повышением уровня экспрессии eGFP. МСЭ – минимальная свободная энергия. Макс. – максимальная; Мин. – минимальная; Ср. – средняя; ШК – штрихкоды; Экспр. – экспрессия (уровень зрелой мРНК eGFP); ρ – коэффициент корреляции Пирсона. * – p-значение <0.05, ** – p-значение <0.01, *** – p-значение <0.001, ***** – p-значение <0.00001.

Библиотека	ρ(МСЭ, экспр.)	Макс. МСЭ	Мин. МСЭ	Ср. МСЭ	Уровень достоверности точного критерия Фишера
17-24	0.1^{*****}	-1	-16.7	-3	0.09
21-28	0.42^{*****}	-6.7	-17	-9.3	$1e^{-31}$
25-32	0.42^{*****}	-10.8	-20.7	-12.5	0
29-36	0.19^{*****}	-1.6	-18.4	-5.5	$2e^{-18}$
33-40	0.17^{*****}	-2.4	-15.9	-3.4	$7e^{-4}$
37-44	0.35^{*****}	-4.8	-24.1	-8.1	$5e^{-213}$
41-48	0.28^{*****}	-11.4	-31.4	-15.1	$3e^{-146}$
45-52	-0.02^{***}	-11.4	-25.4	-13.5	-
49-56	-0.02^{**}	-11.4	-18.4	-12.7	-
49-56 + WT (ШК)	0.01^{*}	-5.7	-29.6	-12.2	-

На следующем этапе анализа мутации были разделены на категории в зависимости от того, имеют ли они уровень экспрессии выше или ниже, чем у исходной последовательности (логарифмическое значение нормированного уровня экспрессии >0 или <0) и имеют ли стабильность вторичной структуры меньше или больше, чем у исходной последовательности (МСЭ > -14.15 ккал/моль или < -14.15 ккал/моль). Данные для каждой библиотеки были занесены в таблицу типа 2×2. С использованием точного критерия Фишера было показано, что понижение стабильности вторичной структуры пре-мРНК в районе +21..+48 н. после СПА статистически достоверно связано с повышением уровня экспрессии eGFP (Таблица 2).

Ранее в ряде работ уже было показано, что повышение стабильности вторичной структуры вблизи СПА в пре-мРНК приводит к снижению уровня экспрессии репортёрного гена (Klasena *et al.*, 1998; Wu and Alwine, 2004). Однако в этих работах авторами было проанализировано не более 20 вариантов последовательностей. В данной работе было проанализировано более 200 тысяч вариантов последовательностей, что позволило впервые обнаружить статистически достоверную взаимосвязь между низкой стабильностью вторичной структуры после СПА и высоким уровнем экспрессии репортёра.

Последовательность ШК оказывает незначительное влияние на уровень экспрессии репортёра. Последовательности ШК могут оказывать влияние на уровень зрелой мРНК eGFP наряду с последовательностями мутаций. В библиотеке 49-56, как было показано ранее, последовательности мутаций не оказывают влияния на уровень экспрессии репортёра и наблюдаемая вариабельность уровней экспрессии репортёра может быть обусловлена именно влиянием последовательности ШК. Для того, чтобы проанализировать влияние ШК, мы объединили данные из библиотеки 49-56 с данными о ШК, ассоциированных с мутацией WT из всех остальных библиотек (Таблица 1, библиотека «49-56 + WT»).

Было обнаружено, что в библиотеке «49-56 + WT» только 0.06% ШК повышают или понижают уровень зрелой мРНК eGFP более чем в 2 раза по сравнению с усреднённым уровнем экспрессии конструкций с мутацией WT (Таблица 1). Также был проведён анализ влияния стабильности вторичной структуры пре-мРНК в районе ШК на уровень экспрессии eGFP. Была проанализирована вторичная структура фрагмента пре-мРНК, включающего в себя 33 н. до ШК, ШК и 10 н. после него. При этом корреляции между МСЭ и уровнем экспрессии eGFP не наблюдалось (Таблица 2). Таким образом, первичная последовательность ШК и его вторичная структура оказывают незначительное влияние на уровень экспрессии репортёра. Для ШК, обеспечивающих высокий уровень экспрессии eGFP, характерно обогащение цитозинами и обеднение гуанинами. С помощью точного критерия Фишера было показано, что ШК с уровнем экспрессии eGFP, превышающим уровень экспрессии конструкций с мутацией WT, были обогащены мотивами ACT, ACG, GAA, GAC, AAGA, TGGA, GACT, CGGA, тогда как ШК с уровнем экспрессии eGFP ниже, чем у исходной последовательности, были обогащены мотивами AGG, GGG, GGC, TAG, AGGG, CAGG, GGCC, TAGG. Интересно, что мотивы, ассоциированные с повышением уровня экспрессии репортёра, отличаются у ШК и мутаций, тогда как мотивы, ассоциированные с понижением уровня экспрессии репортёра, в большинстве своём оказались общими для ШК и мутаций и, вероятно, снижают уровень экспрессии репортёра вне зависимости от своего положения – в 3’-НТО или после сайта полиаденилирования пре-мРНК.

Использование методов классического машинного обучения для обучения моделей, предсказывающих уровень экспрессии репортёра в зависимости от последовательности после СПА или ШК. Для того, чтобы учесть влияние ШК на уровень экспрессии eGFP, нами была получена модель, предсказывающая влияние последовательности ШК на уровень экспрессии репортёра. Она была обучена с использованием методов классического машинного обучения на данных из библиотеки «49-56 + WT». С помощью полученной модели было предсказано влияние всех исследованных в данной работе ШК на уровень экспрессии репортёров.

Также нами с использованием методов классического машинного обучения были обучены 8 моделей, предсказывающих уровень зрелой мРНК на основе последовательности каждого исследованного района после СПА на данных МПРА, рассчитанных MPRAdecoder, из которых были вычтены значения, предсказанные моделью для ШК.

На основе моделей, предсказывающих уровень зрелой мРНК по последовательности района после СПА, была разработана финальная модель, которая суммирует взвешенные предсказания 8 моделей. Веса соответствуют метрике R^2 для предсказаний каждой модели на тестовой выборке, поделённой на сумму R^2 на тестовой выборке для всех 8 моделей. Эта модель была названа DoSIA – Downstream Sequence element Iterative Analyzer. DoSIA позволяет предсказать уровень зрелой мРНК при замене района, расположенного с 3 по 38 н. после сайта полиаденилирования, на любую последовательность соответствующей длины.

Использование DoSIA для предсказания последовательностей DSE, обеспечивающих изменение уровня экспрессии репортёра. Мы использовали DoSIA для предсказания влияния всех возможных однонуклеотидных замен, делеций и инсерций в исходной последовательности после СПА sNRP-1 на уровень экспрессии репортёрного гена eGFP. DoSIA предсказывает повышение экспрессии репортёра при делеции цитозина в позиции +32 п.н. ниже СПА sNRP-1 (+31_AC_A, ΔC). Однако, согласно предсказаниям DoSIA, среди всех возможных однонуклеотидных замен, инсерций и делеций, наибольшее повышение уровня экспрессии репортёра должна обеспечивать замена гуанина на тимин в позиции +33 п.н. ниже СПА sNRP-1 (+33_G_T), а наибольшее снижение – замена тимина на гуанин в позиции +30 п.н. ниже СПА sNRP-1 (+30_T_G).

Мы получили две плазмидные конструкции с этими нуклеотидными заменами ниже СПА sNRP-1 eGFP. Каждой из полученных конструкций были трансфицированы культивируемые клетки HEK293T. Через двое суток после трансфекции мы измерили уровень зрелой мРНК eGFP методом ОТ-кПЦР (Рисунок 3), а уровень белка – методом проточной цитофлуорометрии (Рисунок 4).

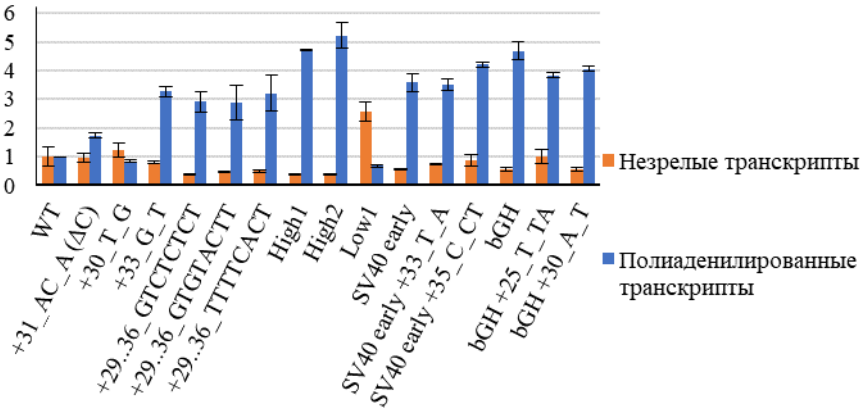

Рисунок 3. Уровень незрелой (непроцессированной) и зрелой (полиаденилированной) мРНК eGFP для индивидуальных мутаций, измеренный методом ОТ-кПЦР. Эксперимент был проведён дважды. Пределы погрешности отражают стандартную ошибку. bGH – бычий гормон роста.

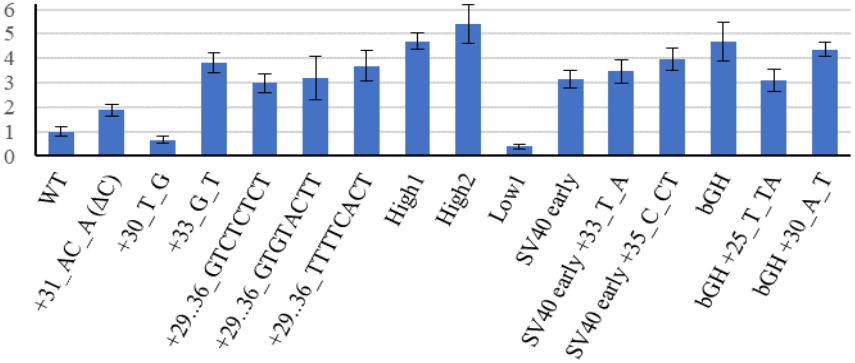

Рисунок 4. Уровень белка eGFP для индивидуальных мутаций, измеренный методом проточной цитофлуорометрии. Эксперимент был проведён трижды. Пределы погрешности отражают стандартную ошибку.

Экспериментальные данные согласуются с предсказаниями DoSIA. Мутация +33_G_T действительно приводит к увеличению уровня зрелой мРНК eGFP в 3.3 раза и уровня белка в 3.8 раза и обеспечивает более высокий уровень зрелой мРНК (p-значение = 0.03, t-критерий Стьюдента) и белка (p-значение = 0.03, t-критерий Стьюдента) по сравнению с мутацией ΔC. Интересно отметить, что мутация +33_G_T обеспечивает такой же уровень зрелой мРНК (p-значение = 0.85, t-критерий Стьюдента) и белка (p-значение = 0.36, t-критерий Стьюдента), как и последовательность терминатора SV40 early. Мутация +30_T_G приводит к снижению уровня зрелой мРНК eGFP в 1.1 раза. Мы предполагаем, что действие этих мутаций связано как с

изменением T-богатости области DSE, так и с их влиянием на стабильность вторичной структуры пре-мРНК после СПА.

Мы также использовали DoSIA для предсказания в последовательностях широко используемых терминаторов транскрипции SV40 early и bGH однонуклеотидных мутаций, оказывающих максимальное влияние на уровень экспрессии репортёрного гена. Плазмидами, несущими эти мутации, были трансфицированы клетки HEK293T. Оказалось, что замены +33_T_A и +35_C_CT в терминаторе SV40 early и +25_T_TA и +30_A_T в терминаторе bGH не привели к статистически достоверным изменениям уровня зрелой мРНК (Рисунок 3) и белка eGFP (Рисунок 4). Таким образом, оптимальные последовательности DSE SV40 early и bGH устойчивы к однонуклеотидным мутациям в отличие от последовательности WT.

Далее мы использовали собственную программу DSEgenerator, которая, основываясь на предсказаниях DoSIA, сгенерировала последовательности DSE, оказывающие наибольшее влияние на уровень экспрессии репортёрного гена. Нами были предсказаны и заклонированы две последовательности High1 и High2, которые, согласно предсказаниям DoSIA, должны повышать уровень экспрессии eGFP, и одна последовательность Low1, которая, согласно предсказаниям DoSIA, должна снижать уровень экспрессии eGFP. Последовательность High1 отличается от WT заменой 27 н., High2 24 н., а Low1 23 н. в области +16..+49 н. после СПА. Последовательности High1 и High2 являются GT-, CT- и T-богатыми. Последовательности High1 и High2 обеспечивают низкую стабильность вторичных структур пре-мРНК после СПА. Последовательность Low1 является T-бедной, GC-богатой и обеспечивает формирование стабильной вторичной структуры в пре-мРНК после СПА.

Плазмидами, несущими эти мутации, были трансфицированы клетки HEK293T. Было показано, что мутации High1 и High2 обеспечивают повышение уровня зрелой мРНК eGFP в 4.7 и 5.2 раза соответственно (Рисунок 3), а уровень белка eGFP в 4.7 и 5.4 раза соответственно (Рисунок 4). Интересно отметить, что мутации High1 и High2 обеспечивают такой же уровень зрелой мРНК (p-значения = 0.77 и 0.81 соответственно, t-критерий Стьюдента) и белка eGFP (p-значения = 0.99 и 0.62 соответственно, t-критерий Стьюдента), как и последовательность терминатора bGH. Мутация Low1 обеспечивает снижение уровня зрелой мРНК eGFP в 1.4 раза (Рисунок 3).

Мы предполагаем, что все исследованные в данной работе мутации DSE влияют на уровень экспрессии репортёрного гена, меняя способность DSE быть связанным комплексом CstF, что, в свою очередь, приводит к изменению эффективности разрезания пре-мРНК. Для того, чтобы проверить это предположение, мы измерили количество непроцессированных транскриптов. Оказалось, что мутация Low1 обеспечивает более высокий уровень

непроцессированной мРНК eGFP по сравнению с мутациями High1 и High2 (p-значения = 0.048 и 0.047 соответственно, t-критерий Стьюдента, Рисунок 3). Между уровнями незрелых и полиаденилированных транскриптов для всех исследованных конструкций наблюдается обратная корреляция Пирсона, равная -0.69.

ЗАКЛЮЧЕНИЕ

В данной работе было показано, что последовательность и вторичная структура РНК в области DSE оказывают большое влияние на уровень зрелой мРНК и белка eGFP.

ВЫВОДЫ

1. Последовательности, расположенные в районе +17..52 п.н. после СПА, влияют на количество зрелой мРНК гена eGFP в культивируемых клетках человека HEK293T. Наибольшее влияние оказывает район +17..40 п.н. после СПА.

2. Для последовательностей, расположенных в районе +17..52 п.н. после СПА, связанных с высоким уровнем зрелой мРНК гена eGFP в культивируемых клетках человека HEK293T, характерна T-обогащённость и наличие мотивов CTC, TCT, TGT, TTG, GTCT, GTGT, TGTC, TCTC. Для последовательностей, расположенных в районе +17..52 п.н. после СПА, связанных с низким уровнем зрелой мРНК гена eGFP в культивируемых клетках человека HEK293T, характерна бедность тимином и наличие мотивов AGG, GGG, GGC, AGGG, CAGG, GGCC, AGGC.

3. Понижение минимальной свободной энергии предсказанной вторичной структуры пре-мРНК после СПА связано со снижением уровня зрелой мРНК репортёрного гена eGFP в культивируемых клетках человека HEK293T.

4. Модификации исходного терминатора транскрипции, содержащие последовательность High1 или High2, в районе +16..+49 н. после СПА, обеспечивают повышение уровня зрелой мРНК и белка eGFP в культивируемых клетках человека HEK293T в 4.7 и 5.2 раза соответственно.

СПИСОК РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Omelina E.S., Ivankin A.V., **Letiagina A.E.**, Pindyurin A.V. Optimized PCR conditions minimizing the formation of chimeric DNA molecules from MPRA plasmid libraries // BMC Genomics. 2019. Vol. 20, № Suppl 7. P. 1–10.

2. **Letiagina A.E.**, Omelina E.S., Ivankin A.V., Pindyurin A.V. MPRAdecoder: Processing of the raw MPRA data with a priori unknown sequences of the region of interest and associated barcodes // Front. Genet. 2021. Vol. 12, № May. P. 1–12.

3. Omelina E.S., **Letiagina A.E.**, Boldyreva L.V., Ogienko A.A., Galimova Yu.A., Yarinich L.A., Pindyurin A.V., Andreyeva E.N. Slight variations in the sequence downstream of the polyadenylation signal significantly increase transgene expression in HEK293T and CHO cells // Int. J. Mol. Sci. 2022. Vol. 23, № 24.