

Федеральное государственное бюджетное учреждение науки
«Институт молекулярной и клеточной биологии СО РАН,
Федеральное государственное автономное образовательное учреждение
высшего образования «Новосибирский национальный исследовательский
государственный университет»

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

Диссертация на соискание ученой степени
кандидата биологических наук

Научный руководитель

к.б.н. Омелина Е.С.

Новосибирск 2025

Оглавление

Список сокращений и условных обозначений	4
Введение	9
<i>Актуальность темы исследования</i>	9
<i>Степень разработанности проблемы</i>	10
<i>Цели и задачи работы</i>	11
<i>Научная новизна</i>	11
<i>Теоретическая и практическая значимость работы</i>	11
<i>Методология и методы исследования</i>	12
<i>Положения, выносимые на защиту</i>	12
<i>Апробация результатов</i>	13
<i>Личный вклад автора</i>	14
<i>Структура и объём работы</i>	15
Обзор литературы	16
<i>Терминация транскрипции</i>	16
<i>Терминация транскрипции у бактерий</i>	16
<i>Rho-независимая терминация</i>	17
<i>Rho-зависимая терминация</i>	18
<i>Терминация транскрипции у эукариот</i>	19
<i>Терминация транскрипции РНК-полимеразы I</i>	20
<i>Терминация транскрипции РНК-полимеразы II</i>	22
Канонический путь процессинга 3'-конца пре-мРНК	24
Процессинг 3'-конца с помощью Star-PAP	41
Альтернативное полиаденилирование	42
Терминация транскрипции РНК-полимеразы II в белок-кодирующих генах	45
Альтернативные механизмы терминации транскрипции РНК-полимеразы II	51
Регуляция терминации транскрипции РНК-полимеразы II	57
<i>Терминация транскрипции РНК-полимеразы III</i>	58
<i>Использование МПРА и машинного обучения для анализа влияния последовательностей цис-элементов на экспрессию генов</i>	59

<i>Дизайн переменных последовательностей</i>	61
<i>Анализ молекулярного фенотипа репортёров</i>	62
<i>Анализ данных высокопроизводительного секвенирования</i>	64
<i>Дизайн последовательностей с использованием моделей, обученных на данных МПРА</i>	65
<i>Исследование процессинга 3'-конца пре-мРНК с помощью МПРА и нейронных сетей</i>	67
<i>Краткая история исследований DSE</i>	70
Материалы и методы	74
Результаты	95
<i>Проведение МПРА</i>	95
<i>Дизайн библиотек для МПРА</i>	95
<i>Оптимизация условий ПЦР позволяет снизить долю химерных продуктов</i>	96
<i>Подготовка образцов для МПРА</i>	106
<i>Анализ данных МПРА</i>	107
<i>Наибольшее влияние на уровень зрелой мРНК оказывают последовательности, расположенные в районе +17..+40 н. после СПА</i>	110
<i>Низкая стабильность вторичной структуры после СПА коррелирует с высоким уровнем экспрессии репортёра</i>	114
<i>Последовательность ШК оказывает незначительное влияние на уровень экспрессии репортёра</i>	117
<i>Использование методов классического машинного обучения для обучения моделей, предсказывающих уровень экспрессии репортёра в зависимости от последовательности после СПА или ШК</i>	119
<i>Сравнение полученной нами модели DoSIA с уже известной APARENT2</i>	121
<i>Использование DoSIA для предсказания последовательностей DSE, обеспечивающих изменение уровня экспрессии репортёра</i>	124
Заключение	138
Выводы	143
Благодарности	144
Список литературы	145
Приложение 1	192
Приложение 2	196

Список сокращений и условных обозначений

АПА – альтернативное полиаденилирование
БСА – бычий сывороточный альбумин
ВПГ-1 – Вирус простого герпеса 1 типа
гДНК – геномная ДНК
днкРНК – длинная не кодирующая РНК
кПЦР – количественная полимеразная цепная реакция
ЛКП – локус количественных признаков
МЕ – международная единица
МПРА – массовый параллельный репортёрный анализ
МСЭ – минимальная свободная энергия
мякРНК – малая ядрышковая РНК
мяРНК – малая ядерная РНК
мяРНП – малый ядерный рибонуклеопротеин
нкРНК – не кодирующая РНК
НТО – нетранслируемая область
ОТ – обратная транскрипция
оцРНК – одноцепочечная РНК
пДНК – плазмидная ДНК
пре-мРНК – предшественник матричной РНК
рДНК – рибосомная ДНК
рРНК – рибосомная РНК
СПА – сигнал полиаденилирования
РНКазы - рибонуклеазы
т.п.н. – тысяча пар нуклеотидов
тРНК – транспортная РНК
ШК – штрихкод
эПЦР – эмульсионная ПЦР

AmpR – ampicillin resistance, устойчивость к ампициллину

APARENT – APA REgression NeT, регрессионная сеть АПА

APT – associated with Pta1, связанный с Pta1

ARS – autonomously replicating sequence, автономно реплицирующаяся последовательность

bGH – bovine growth hormone, бычий гормон роста

CDK – cyclin-dependent kinase, циклин-зависимая киназа

CF – cleavage factor, фактор разрезания

CHO – chinese hamster ovary, яичник китайского хомячка

CLP1 – cleavage factor polyribonucleotide kinase subunit 1, полирибонуклеотидкиназа, субъединица 1 фактора разрезания

CoTC – cotranscriptional cleavage, котранскрипционное разрезание

CPF – cleavage and polyadenylation factor, фактор разрезания и полиаденилирования

CPSF – cleavage and polyadenylation specificity factor, фактор специфичности разрезания и полиаденилирования

CstF – cleavage stimulation factor, фактор, стимулирующий разрезание

CTD – carboxy-terminal domain, С-концевой домен

DMEM – Dulbecco's modified Eagle medium, среда Игла, модифицированная Дульбекком

DoSIA – Downstream Sequence element Iterative Analyzer

DSE – downstream sequence element, элемент последовательности, расположенный ниже СПА

EDTA – ethylenediaminetetraacetic acid, этилендиаминтетрауксусная кислота

eGFP – enhanced green fluorescent protein, улучшенный зелёный флуоресцентный белок

ETS – external transcribed spacer, внешний транскрибируемый спейсер

FACS – fluorescence-activated cell sorting, сортировка клеток, активированная флуоресценцией

FBS – fetal bovine serum, фетальная бычья сыворотка

FLASH – FLICE-associated huge protein, огромный белок, ассоциированный с FLICE

HCC – histone pre-mRNA cleavage complex, комплекс, разрезающий пре-мРНК гистонов

HDE – histone downstream element, элемент последовательности гистонов, расположенный ниже консервативной шпильки

HEK – human embryonic kidney, почки эмбриона человека

hCMV – human cytomegalovirus, цитомегаловирус человека

hPGK – human phosphoglycerate kinase 1, Фосфоглицераткиназа 1 человека

IgM – immunoglobulin M, иммуноглобулин класса M

IGS – intergenic spacer, межгенный спейсер

IMDM – Iscove's Modified Dulbecco's Medium, среда Дульбекко, модифицированная Исков

INT – Integrator

ITS – internal transcribed spacer, внутренний транскрибируемый спейсер

K – основания G или T

Mpe1 – mutant Pcf11 extragenic suppressor, экстрагенный супрессор мутантного Pcf11

MSE – mean squared error, среднеквадратичная ошибка

N – основания A, C, G или T

NGS – next generation sequencing, секвенирование нового поколения

NRT – no reverse transcriptase, без обратной транскриптазы

Nsi1 – nontranscribed spacer silencing protein 1 (белок сайленсинга 1 нетранскрибируемого спейсера 1)

NTC – no template control, контроль без матрицы

NTD – N-terminal domain, N-концевой домен

OBR – origin of bidirectional DNA replication, ориджин двунаправленной репликации

PABPC – cytosolic poly(A)-binding protein, цитозольный поли(А)-связывающий белок

PABPN1 – polyadenylate-binding nuclear protein 1, ядерный поли(А)-связывающий белок 1

PAP – poly(A) polymerase, поли(А)-полимераза

PBS – phosphate buffered saline, натрий-фосфатный буфер

Pfs – polyadenylation factor subunit, субъединица фактора полиаденилирования

piPDK – PDK, взаимодействующая с Piwi

PNUTS – phosphatase 1 nuclear targeting subunit, субъединица фосфатазы 1, направляющая её в ядро

PP – protein phosphatase, белковая фосфатаза

PROMPT – promoter upstream transcript, транскрипт последовательности выше промотора

PSF – polyadenylation specificity factor, фактор специфичности полиаденилирования

PSR – pre-mRNA sensing region, участок, чувствующий пре-мРНК

PTRF – RNA polymerase I and transcript release factor, фактор высвобождения РНК-полимеразы I и транскрипта

R – основания A или G

R² – коэффициент детерминации

RBBP6 – Retinoblastoma-binding protein 6, ретинобластома-связывающий белок 6

RFB – replication fork barrier, барьер для репликационной вилки

Rnt1 – RNase III-like endonuclease, эндонуклеаза, подобная РНКазе III

RRM – RNA recognition motif, мотив, распознающий РНК

S – основания C или G

SETD – SET domain containing, содержащая домен SET

sNRP-1 – soluble neuropilin-1, растворимый нейропилин 1

Star-PAP – speckle targeted PIP5K1A-regulated poly(A) polymerase, направляемая в спеклы PAP, регулируемая фосфатидинозитол-4-фосфат-5-киназой типа 1 α

TR – terminal repeat, терминальный повтор

TTF-I – transcription termination factor for RNA polymerase I, фактор терминации транскрипции для РНК-полимеразы I

TUT – terminal uridylyl transferase, терминальная уридилтрансфераза

UBL – ubiquitin-like domain, убиквитин-подобный домен

USE – upstream sequence element, элемент последовательности, расположенный выше СПА

V – основания A, C или G

W – основания A или T

WT – wild type, дикий тип

XRN – 5'-3' exoribonuclease, 5'-3' экзорибонуклеаза

Y – основания C или T

ρ – коэффициент корреляции Пирсона

σ – стандартное отклонение

Введение

Актуальность темы 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 [1]. 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 (Приложение 1). 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 [2–6].

Ранее сотрудниками нашей лаборатории было показано, что делеция одного цитозина в позиции +32 п.н. ниже СПА sNRP-1 (Δ C) приводит к двукратному повышению уровня транскриптов репортёрного гена eGFP в культивируемых эмбриональных стволовых клетках мыши mESCs, мышинных клетках 3T3 и клетках человека HEK293T [7,8]. Таким образом, можно предположить, что последовательность Δ C 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 [8]. 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, расширяя научные представления о влиянии последовательности, расположенной после сайта полиаденилирования, и её вторичной структуры на уровень зрелой мРНК и эффективность процессинга пре-мРНК репортёрного гена. В рамках работы были разработаны подходы, позволяющие использовать DSE в качестве области, регулирующей уровень эписомальной экспрессии целевого гена, что может иметь практическое применение в 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 после 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 были представлены и обсуждались на нескольких международных и всероссийских научных конференциях:

1. Пиндюрин А.В., Болдырева Л.В., Иванкин А.В., **Лещенко А.Е.** и др. Вклад последовательностей ДНК, располагающихся в районе терминации транскрипции, в уровень экспрессии генов // Хромосома. Новосибирск, 2018. С. 155.
2. Omelina E., Ivankin A., **Leshchenko A.** и др. Two-step emulsion PCR to prevent formation of the chimeric molecules // Биоинформатика регуляции и структуры генома\системная биология (BGRS\SB-2018). Новосибирск, 2018. С. 64.
3. Boldyreva L.V., **Leshchenko A.E.** и др. Large-scale screening for gene expression regulatory DNA motifs located downstream transcription termination sites // Chromosomes and Mitosis. Новосибирск, 2019. С. 8-9.
4. Пиндюрин А.В., Болдырева Л.В., Иванкин А.В., **Летягина А.Е.** и др. Широкомасштабный поиск нуклеотидных мотивов, регулирующих процесс терминации транскрипции у млекопитающих // VII съезд ВОГиС. Санкт-Петербург, 2019. С. 471.
5. Болдырева Л.В., Омелина Е.С., **Летягина А.Е.** и др. Поиск нуклеотидных мотивов, регулирующих процесс терминации транскрипции // II Объединённый научный форум. VI Съезд физиологов СНГ. VI Съезд биохимиков России. IX Российский симпозиум "Белки и пептиды". Москва, 2019. С. 11.
6. Andreyeva E.N., **Letiagina A.E.** и др. Multiplex analyses of the effect of transcription termination on gene activity in eukaryotes // Хромосома. Новосибирск, 2023. С. 7.
7. **Летягина А.Е.** и др. Небольшие изменения последовательности ДНК, расположенной после сайта полиаденилирования, влияют на экспрессию

трансгена в клетках линий НЕК293Т и СНО // Хромосома. Новосибирск, 2023. С. 135-136.

8. **Летягина А.Е.** и др. Разработка модели, способной предсказать уровень зрелой мРНК в культивируемых клетках млекопитающих на основе последовательности DSE терминатора транскрипции // VIII съезд ВОГиС. Саратов, 2024. С. 386.
9. **Летягина А.Е.** Рациональный дизайн терминаторов транскрипции РНК-полимеразы II для повышения наработки белков в клетках млекопитающих // НБНМ. Самара, 2024. С. 194.

По теме диссертации были опубликованы 3 работы в рецензируемых научных журналах, входящих в международные базы цитирования (WoS, Scopus):

1. Omelina E.S., Ivankin A.V., **Letiagina A.E.** *et al.* 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.** *et al.* 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.** *et al.* 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.

Во второй публикации Летягина А.Е. является первым автором, в третьей – со-первым автором.

Личный вклад автора

Большая часть результатов, включенных в диссертацию, были получены автором самостоятельно. Летягина А.Е. принимала непосредственное участие в планировании экспериментов, обсуждении и интерпретации результатов, подготовке публикаций. Автор работы самостоятельно провела подготовку

образцов для секвенирования библиотек МПРА, написала программу для анализа данных МПРА и проанализировала данные МПРА. Автор работы обучила модель, предсказывающую уровень экспрессии репортёрного гена на основе последовательности после СПА. С помощью этой модели она предсказала и клонировала последовательности, влияющие на уровень экспрессии репортёра. Автор измерила уровень зрелой мРНК и белка репортёрного гена, который обеспечивают полученные конструкции при эписомальной экспрессии в культивируемых клетках НЕК293Т.

Дизайн библиотек для МПРА был разработан предыдущим заведующим лабораторией клеточного деления ИМКБ СО РАН Пиндюриным А.В. и с.н.с. лаборатории клеточного деления ИМКБ СО РАН Болдыревой Л.В. Плазмидные библиотеки для МПРА были получены н.с. лаборатории клеточного деления ИМКБ СО РАН Яринич Л.А. Подбор оптимальных условий ПЦР для подготовки образцов картирования был проведён автором совместно с.н.с. лаборатории клеточного деления ИМКБ СО РАН Омелиной Е.С. Индексы для образцов картирования, нормирования и экспрессии были разработаны н.с. лаборатории клеточного деления ИМКБ СО РАН Иванкиным А.В. Запуск Illumina MiSeq был проведён автором и Омелиной Е.С. совместно с с.н.с. лаборатории геномики ИМКБ СО РАН Лактионовым П.П. и н.с. лаборатории геномики ИМКБ СО РАН Максимовым Д.А. Нештрихкодированные плазмидные конструкции с терминаторами SV40 early и bGH были получены с.н.с. лаборатории клеточного деления ИМКБ СО РАН Андреевой Е.Н.

Структура и объём работы

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.

Обзор литературы

Терминация транскрипции

Регуляция терминации транскрипции генов традиционно привлекает гораздо меньшее внимание исследователей чем процесс инициации транскрипции. Однако процесс терминации транскрипции является жизненно важным для клетки. Он предотвращает нарушение транскрипции расположенных ниже генов [9] и, предположительно, способствует возвращению РНК-полимеразы II на новые раунды транскрипции [10]. Также было показано, что эффективная терминация транскрипции увеличивает количество мРНК и белка за счёт повышения эффективности сплайсинга и предотвращения деградации образующегося транскрипта [2]. Известно, что неэффективная терминация транскрипции некоторых генов может приводить к образованию антисмысловых РНК к транскриптам расположенных ниже генов. Такие антисмысловые РНК могут индуцировать формирование гетерохроматина и ограничивать экспрессию генов [11,12]. Таким образом, терминация транскрипции влияет на многие аспекты экспрессии генов у эукариот и имеет огромный регуляторный потенциал.

Механизмы терминации транскрипции очень разнообразны. Ниже будут рассмотрены пути терминации транскрипции для некоторых наиболее изученных РНК-полимераз с акцентом на РНК-полимеразу II, транскрибирующую все белок-кодирующие гены эукариот.

Терминация транскрипции у бактерий

Своевременная терминация транскрипции очень важна для бактериальной клетки. Нарушение терминации транскрипции у бактерий может привести к включению дополнительных цистронов в мРНК и, следовательно, нарушить регуляцию синтеза некоторых белков, что может иметь нежелательные последствия для клетки. Бактерии используют два

различных механизма терминации транскрипции: Rho-независимая (внутренняя) и Rho-зависимая терминация (Рисунок 1).

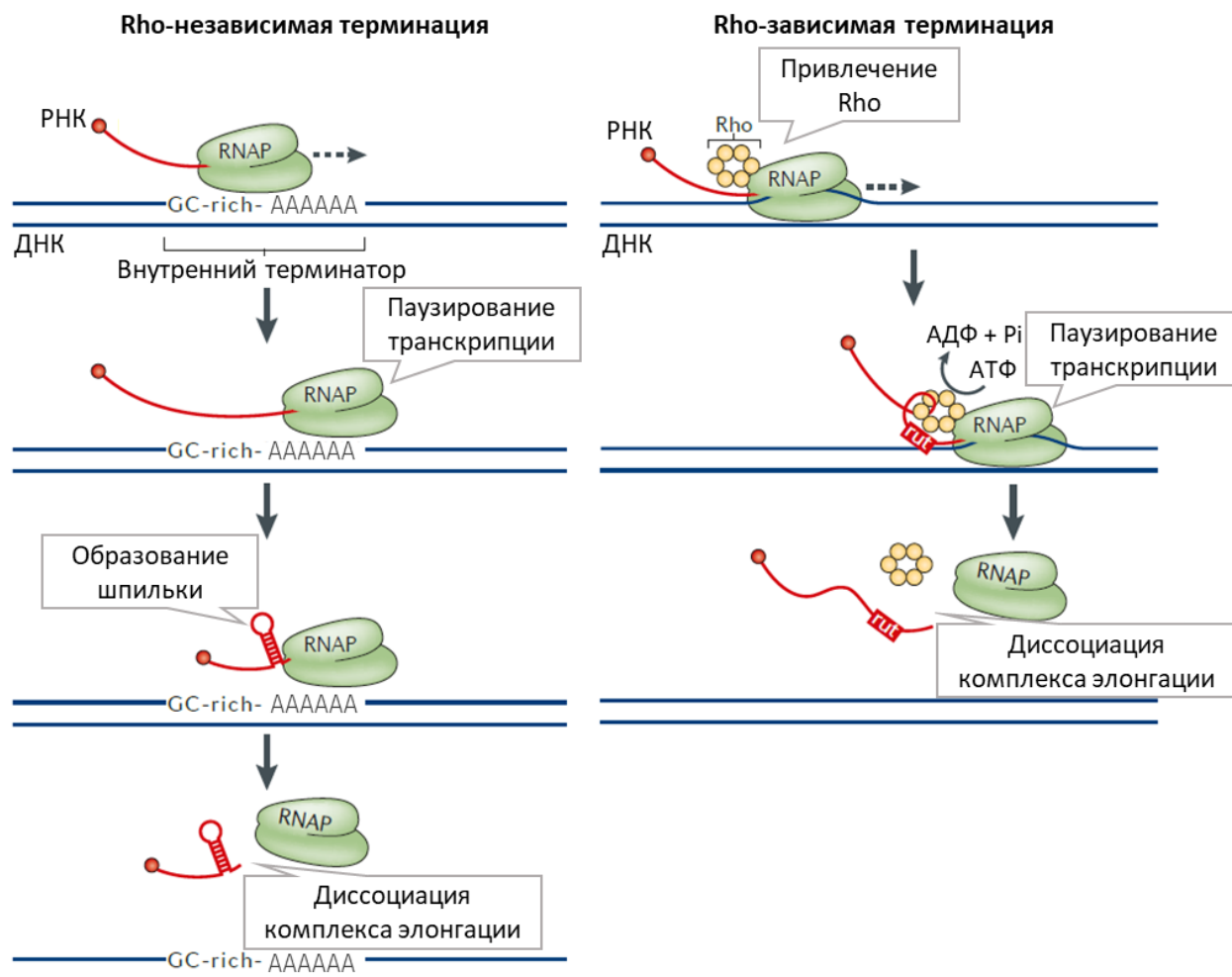


Рисунок 1. Механизмы терминации транскрипции у бактерий в соответствии с [13]. RNAP – РНК-полимераза; rut – сайт связывания Rho-фактора.

Rho-независимая терминация

Rho-независимая терминация не нуждается в привлечении специфических белковых факторов и обусловлена последовательностью ДНК, которая, при транскрипции, приводит к образованию GC-богатой шпильки в образующейся РНК (Рисунок 1, левая панель). За GC-богатой шпилькой расположена А-богатая последовательность. Вместе они образуют внутренний терминатор транскрипции. В процессе транскрипции А-богатой последовательности бактериальная РНК-полимераза останавливается. Гибрид U-богатой РНК и А-богатой ДНК обладает наименьшей термодинамической

стабильностью [14,15]. Низкая стабильность РНК:ДНК гибрида в составе тройного комплекса может вызывать обратный ход РНК-полимеразы. При этом GC-богатая РНК-шпилька съезжает к каналу выхода 5'-конца РНК на РНК-полимеразе бактерий. Это индуцирует диссоциацию комплекса элонгации [16]. Внутренние терминаторы расположены на 3'-конце белок-кодирующих генов или внутри генов, где они выполняют функцию аттенуаторов транскрипции [13].

Rho-зависимая терминация

Rho-зависимая терминация играет роль в формировании 3'-конца мРНК, а также в других процессах, таких, как сайленсинг экспрессии чужеродной ДНК [17], сопряжение транскрипции и трансляции [18] и супрессия транскрипции некодирующей антисмысловой РНК [19]. Rho-фактор – это АТФ-зависимая хеликаза, состоящая из шести идентичных субъединиц, формирующих структуру в форме кольца (Рисунок 1, правая панель). Он взаимодействует с rut (Rho utilization)-сайтом в образующейся РНК. rut-сайты представляют собой С-богатые G-бедные последовательности длиной около 80 нуклеотидов, которые содержат несколько стабильных вторичных структур [20]. Rho-фактор транслоцирует РНК сквозь центр кольца в направлении 5'→3'. РНК-полимераза останавливается в сайте паузирования $G_{-10}Y_{-1}G_{+1}$ [21]. Этот сайт также встречается в некоторых Rho-независимых терминаторах [22]. Rho-фактор «догоняет» паузированную РНК-полимеразу, что, в конечном счёте, приводит к диссоциации комплекса элонгации [23]. Взаимодействие Rho-фактора с rut-сайтом является ключевым для терминации, поэтому активность Rho-зависимых терминаторов часто регулируется специфическими факторами, такими, как РНК-связывающий белок CsrA [24], или рибопереклювателями – альтернативными структурами РНК, экспонирующими или изолирующими rut-сайты в зависимости от различных физиологических условий [25]. Трансляционная машина также может блокировать доступ Rho-фактора к rut-сайтам. Регуляция экспрессии

Rho контролируется Rho-зависимой ранней терминацией, которая приводит к образованию нефункционального транскрипта [26].

У *Escherichia coli* (*E. coli*) около 20% из 4500 генов, кодирующих мРНК, малые РНК и транспортные РНК (тРНК), используют Rho-зависимый механизм терминации [17,27]. Также важной функцией Rho-фактора является репрессия антисмысловой транскрипции [19,28] и разрешение R-петель [29,30]. Нарушение функции Rho-фактора у *E. coli* приводит к образованию двуцепочечных разрывов [31]. Возможно, это связано с конфликтом транскрипции и репликации.

Таким образом, контакт РНК-полимеразы с РНК-шпилькой или Rho-фактором, по-видимому, запускает конформационные изменения в РНК-полимеразе, что приводит к её переходу из элонгирующего состояния в терминирующее [32].

Терминация транскрипции у эукариот

У эукариот, в отличие от бактерий, есть 3 основных РНК-полимеразы, которые выполняют различные функции и используют отличающиеся механизмы терминации транскрипции. Основными объектами для изучения терминации транскрипции являются дрожжи и млекопитающие.

РНК-полимераза I транскрибирует рибосомные РНК (рРНК) 5.8S, 18S и 25S/28S (Рисунок 2А). Транскрипты РНК-полимеразы I представлены в клетке в очень большом количестве.

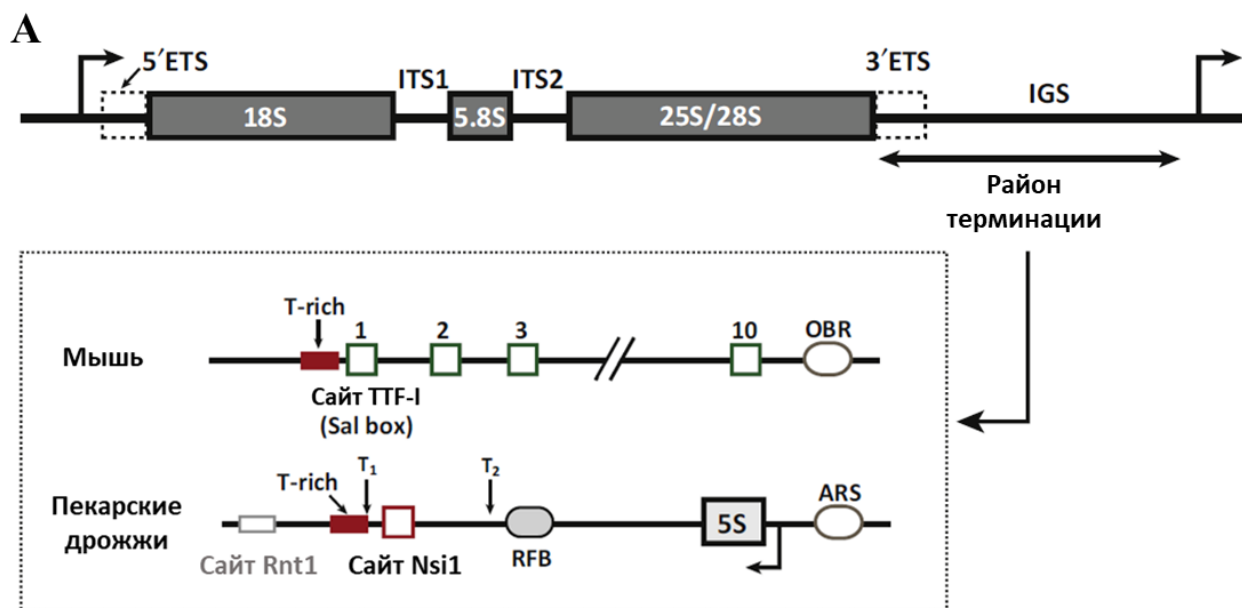
РНК-полимераза II транскрибирует мРНК и разнообразные типы некодирующей РНК (нкРНК). Эти нкРНК могут быть стабильными и представленными в большом числе копий, например малые ядерные и малые ядрышковые РНК (мяРНК и мякРНК соответственно), или представленными в малом количестве и быстро деградировать, как, например, длинные некодирующие РНК (днкРНК). У млекопитающих транскрипты РНК-полимеразы II могут варьировать от нескольких сотен до 10^8 п.н. [33].

РНК-полимераза III транскрибирует тРНК, 5S рРНК и некоторые нкРНК. Транскрипты РНК-полимеразы III как правило короткие (менее 400 п.н.), структурированные и представлены большим числом копий.

Терминация транскрипции РНК-полимеразы I

Рибосомная ДНК организована в множественные тандемные повторы (150-200 копий у дрожжей и 400 копий у человека) и транскрибируется в ядрышке. Каждый повтор содержит последовательность, соответствующую различным зрелым рРНК (5.8S, 18S и 25S у дрожжей и растений и 28S у животных), фланкированным внутренними (ITS1, ITS2) и внешними (ETS) транскрибируемыми спейсерами (Рисунок 2А). Различные повторы разделены межгенным спейсером (IGS). Терминация транскрипции РНК-полимеразы I происходит в IGS, который содержит один или несколько сайтов узнавания ДНК-связывающих белков семейства Myb (от myeloblastosis – миелоидный лейкоз). Перед сайтом связывания белков семейства Myb часто расположена Т-богатая последовательность. Как и в случае Rho-независимой терминации у бактерий Т-богатая последовательность может дестабилизировать ДНК:РНК гибрид в составе тройного комплекса. Большинство исследований сходятся в том, что транскрибирующая РНК-полимераза I блокируется белком семейства Myb и после высвобождается при участии других факторов [34].

У млекопитающих транскрипцию РНК-полимеразы I блокирует белок семейства Myb TTF-I (фактор терминации транскрипции для РНК-полимеразы I). TTF-I связывает последовательность длиной 18 п.н., известную как Sal box. Напрямую с TTF-I связывается PTRF – фактор высвобождения РНК-полимеразы I и транскрипта (Рисунок 2Б). PTRF обеспечивает диссоциацию комплекса элонгации РНК-полимеразы I. У животных IGS содержит несколько тандемно повторённых сайтов связывания TTF-I (до 10 у мыши) [34]. Предполагается, что это связано с низкой эффективностью терминатора транскрипции РНК-полимеразы I.



Б

Терминация РНК-полимеразы I у мыши

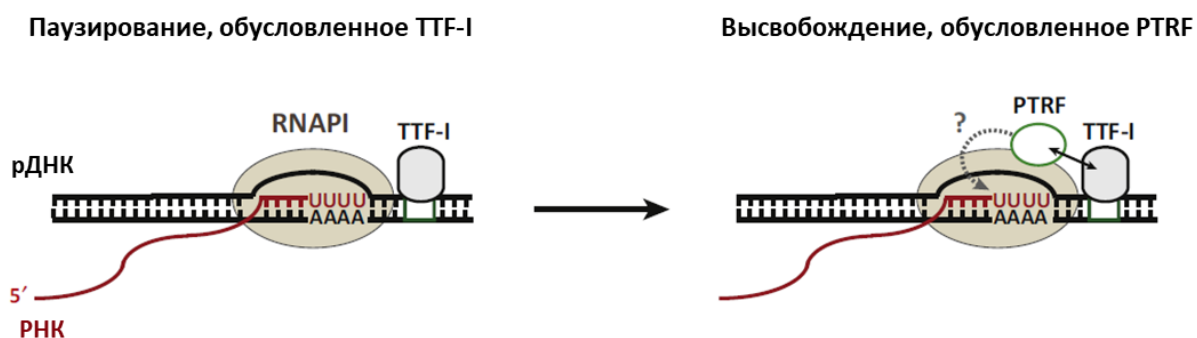


Рисунок 2. Схема терминация транскрипции РНК-полимеразы I в соответствии с [22]. **А)** Структура локуса рДНК. Организация района терминации показана детально для млекопитающих и дрожжей. T₁ и T₂ – альтернативные сайты терминации в сайте связывания Nsi1 и RFB (барьер для репликационной вилки) соответственно. Зелёные прямоугольники в районе терминации млекопитающих обозначают альтернативные и избыточные сайты терминации. **Б)** Модель терминации транскрипции у мыши. ARS – автономно реплицирующаяся последовательность; ETS – внешний транскрибируемый спейсер; IGS – межгенный спейсер; ITS – внутренний транскрибируемый спейсер; Nsi1 – белок сайленсинга 1 нетранскрибируемого спейсера 1; OBR – ориджин двунаправленной репликации; PTRF – фактор высвобождения РНК-полимеразы I и транскрипта; RFB – барьер репликационной вилки; RNAPI – РНК-полимераза I; Rnt1 – эндонуклеаза, подобная РНКазе III; TTF-I – фактор терминации транскрипции для РНК-полимеразы I.

У пекарских дрожжей (*Saccharomyces cerevisiae*) транскрипцию РНК-полимеразы I блокирует белок семейства Myb Nsi1 (белок сайленсинга 1 нетранскрибируемого спейсера 1) [35,36]. У дрожжей IGS содержит только один сайт связывания Nsi1. PTRF-подобный фактор у дрожжей не обнаружен, однако ряд исследований указывает на то, что он может обеспечивать диссоциацию остановленного комплекса элонгации транскрипции РНК-полимеразы I [37,38]. Белок Nsi1 не является жизненно важным, что предполагает наличие у дрожжей дополнительных механизмов терминации транскрипции РНК-полимеразы I. Прямое наблюдение транскрипции РНК-полимеразы I у дрожжей методами электронной микроскопии показало, что в ~90% случаев транскрипция РНК-полимеразы I терминируется непосредственно выше сайта связывания Nsi1, тогда как в оставшихся 10% случаев терминация транскрипции происходит в других районах, предположительно, благодаря альтернативным механизмам терминации [39].

Терминация транскрипции РНК-полимеразы II

Механизмы терминации транскрипции РНК-полимеразы II существенно отличаются от механизмов терминации транскрипции других рассмотренных здесь РНК-полимераз. Во-первых, при транскрипции белок-кодирующих генов диссоциации тройного комплекса элонгирующей РНК-полимеразы II, как правило, предшествует процессинг 3'-конца молекулы пре-мРНК, заключающийся в разрезании и полиаденилировании транскрипта. Полиаденилирование играет важную роль в обеспечении экспорта мРНК из ядра, её стабильности и трансляции [40–46]. Во-вторых, РНК-полимераза II может быть очень процессивной. Её транскрипты у млекопитающих могут превышать 100 т.п.н. [33]. Таким образом, нужны достаточно эффективные терминирующие механизмы, чтобы остановить транскрипцию РНК-полимеразы II. В-третьих, механизмы терминации транскрипции РНК-полимеразы II отличаются разнообразием и большой сложностью. Это может

быть связано с регуляторной ролью терминации транскрипции РНК-полимеразы II.

Субъединица Rpb1 РНК-полимеразы II имеет уникальный С-терминальный домен (CTD), который отделён от основного глобулярного фермента [47]. CTD играет важную роль в координации котранскрипционного процессинга РНК – кэппирования, сплайсинга, разрезания 3'-конца и полиаденилирования. CTD состоит из tandemных повторов гептапептида Tyr1- Ser2- Pro3- Thr4- Ser5- Pro6- Ser7 (26 повторов у дрожжей и 52 у позвоночных) [48–52]. CTD используется как платформа для связывания многих факторов, вовлечённых во все стадии транскрипции и связанные с ней процессы (кэпирование, сплайсинг, полиаденилирование и другие этапы процессинга РНК) [49,53]. На разных стадиях транскрипции определённые аминокислотные остатки CTD подвергаются пост-трансляционным модификациям [54–58] (Рисунок 3). Наибольшее влияние на процесс терминации транскрипции оказывает фосфорилирование остатков Tyr1, Ser2, Thr4, Ser5 и Ser7, меняющее специфичность связывания факторов терминации.

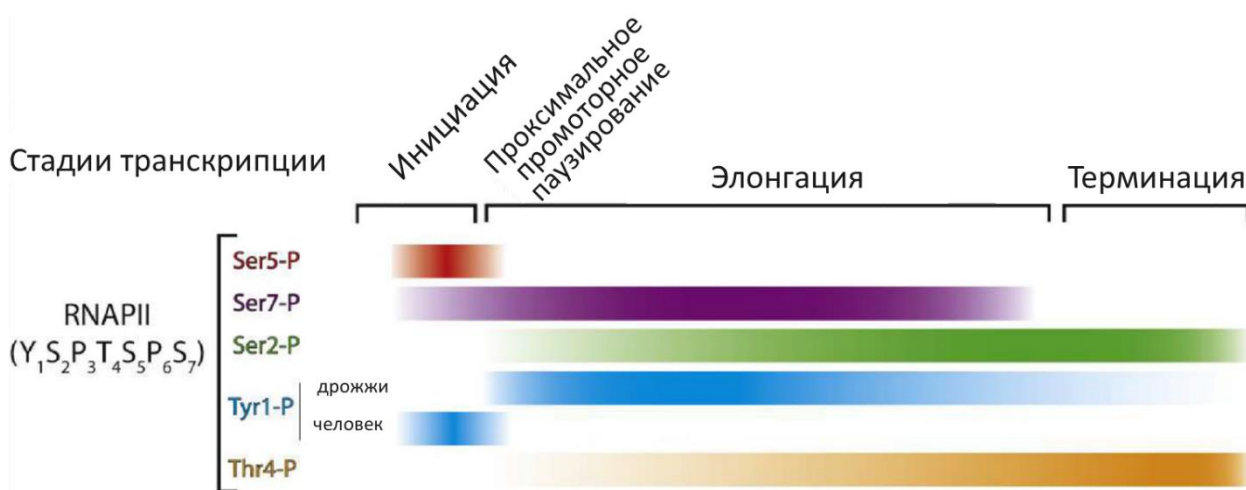


Рисунок 3. Паттерн фосфорилирования различных аминокислотных остатков домена CTD РНК-полимеразы II в ходе цикла транскрипции в соответствии с [59]. RNAPII – CTD РНК-полимеразы II.

Канонический путь процессинга 3'-конца пре-мРНК

Для начала рассмотрим канонический путь процессинга 3'-конца пре-мРНК большинства эукариотических генов и механизмы последующей терминации транскрипции РНК-полимеразы II. В процессинг 3'-конца пре-мРНК вовлечено множество цис-элементов и более 20 белков в клетках дрожжей [40] и 80 белков в клетках человека [60] (Рисунок 4, Таблица 1). Большинство факторов процессинга 3'-конца пре-мРНК кодируются жизненно важными генами [61].

В состав канонического комплекса процессинга 3'-конца пре-мРНК у человека входят следующие мультисубъединичные комплексы: CPSF (который некоторые исследователи подразделяют на комплексы mPSF и mCF [62]), CstF, CFIm, CFIIIm, а также ряд других белков (Рисунок 4) [60]. Комплекс mPSF распознаёт СПА (PAS) и рекрутирует поли(А)-полимеразу (РАР), mCF разрезает пре-мРНК в сайте разрезания [63,64]. CstF распознаёт элемент последовательности, расположенный ниже СПА (DSE) [65]. CFIm связывает мотив UGUA, известный как элемент последовательности, расположенный в пре-мРНК выше СПА (USE) [66]. Роль CFIIIm не вполне ясна.

У дрожжей комплекс процессинга 3'-конца пре-мРНК имеет много общего с человеческим [42]. Дрожжевой комплекс формируется из трёх основных факторов: фактор разрезания и полиаденилирования (CPF), фактор разрезания IA (CFIA) и фактор разрезания IB (CFIB) [40,41]. Все белки CPF дрожжей гомологичны белкам CPSF млекопитающих, а белки CFIA дрожжей имеют гомологов в комплексах CstF и CFII млекопитающих (Таблица 1). Комплекс CFIm обнаружен только у многоклеточных животных и не имеет аналогов у дрожжей [42]. CFIm также является единственным комплексом, не обязательным для процессинга 3'-конца пре-мРНК [67].

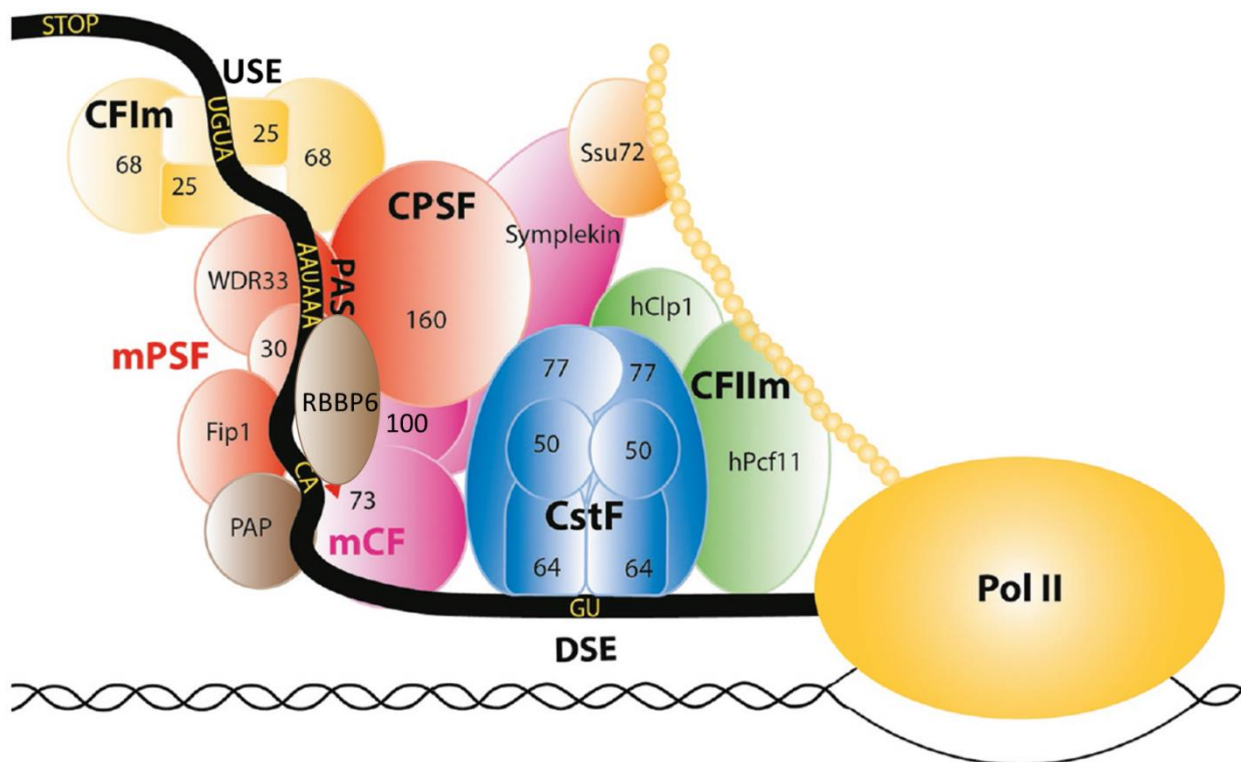


Рисунок 4. Канонический комплекс процессинга 3'-конца пре-мРНК у человека по [62] с изменениями. CFIm – фактор разрезания I млекопитающих; CFIIIm – фактор разрезания II млекопитающих; CPSF – фактор специфичности разрезания и полиаденилирования; CstF – фактор, стимулирующий разрезание; DSE – элемент последовательности, расположенный ниже СПА; hClp1 – полирибонуклеотидкиназа человека, субъединица 1 фактора разрезания; mCF – фактор разрезания млекопитающих; mPSF – фактор специфичности полиаденилирования млекопитающих; PAP – поли(А)полимераза; PAS – сигнал полиаденилирования (СПА); Pol II – РНК-полимераза II; RBBP6 – ретинобластома-связывающий белок 6; USE – элемент последовательности, расположенный выше СПА. Комплекс mPSF окрашен красным, mCF пурпурным, CstF голубым, CFIm жёлтым и CFIIIm зелёным. CTD РНК-полимеразы II изображён в виде цепочки из жёлтых звеньев.

Таблица 1 (начало). Белки, принимающие участие в процессинге 3'-конца пре-мРНК и их дрожжевые ортологи в соответствии с [42,67,68]. Цвет ячейки обозначает соответствие белка определённому мультисубъединичному комплексу, представленному на Рисунке 4.

Мультисубъединичный комплекс млекопитающих	Белок млекопитающих	Роль	Дрожжевой ортолог	Мультисубъединичный комплекс дрожжей
CPSF (mPSF)	CPSF160 (CPSF1)	Скэффолд	Yhh1 (Cft1)	CPF (Кор, модуль PAP)
	WDR33	Связывает РНК (СПА)	Pfs2	
	CPSF30 (CPSF4)	Связывает РНК (СПА), рекрутирует hFip1	Yth1	
	hFip1	Связывает РНК и PAP	Fip1	
CPSF	PAP	Поли(А)-полимераза	Pap1	CPF (Кор, нуклеазный модуль)
	RBBP6	Активирует разрезание, связывает РНК	Mpe1	
CPSF (mCF)	CPSF100 (CPSF2)	Псевдоэндонуклеаза, сборка	Ydh1 (Cft2)	
	CPSF73 (CPSF3)	Эндонуклеаза	Ysh1 (Brr5)	
	Symplekin	Скэффолд	Pta1	CPF (фосфатазный модуль)/APT
	Ssu72	Фосфатаза	Ssu72	
Фосфатазный комплекс	WDR82	Терминация транскрипции	Swd2	
	PP1	Протеиновая фосфатаза	Glc7	
	PNUTS	Скэффолд, активатор PP1/Glc7	Ref2	
	-	Скэффолд	Pti1	
	Tox4	Связывание ДНК	-	

Таблица 1 (продолжение).

Мультисубъединичный комплекс млекопитающих	Белок млекопитающих	Роль	Дрожжевой ортолог	Мультисубъединичный комплекс дрожжей
	-	Антагонист Ysh1	Syc1	APT
CstF	CstF77 (CstF3)	Связывает CPSF160/Cft1	Rna14	CFIA
	CstF64 (CstF2) / CstF64τ	Связывает РНК (DSE)	Rna15	
	CstF50 (CstF1)	Стабилизирует комплекс	-	
CFIIm	hPcf11	Связывает CTD РНК-полимеразы II	Pcf11	CF IA
	hClp1	РНК-киназа	Clp1	
CFIm	CFIm25 (CPSF5, NUTD21)	Связывает РНК (USE)	-	
	CFIm68 (CPSF6)	Связывает hFip1	-	
	CFIm59 (CPSF7)	Связывает hFip1	-	
	-	Обеспечивает точность разрезания	Hrp1	CF IB
	PABPN1	Связывает поли(А)-хвост в ядре	-	
	-	Связывает поли(А)-хвост в ядре	Pab1	
	ZC3H14	Связывает поли(А)-хвост в ядре	Nab2	

Ключевым цис-элементом, обеспечивающим сборку канонического комплекса процессинга 3'-конца пре-мРНК, является СПА (Рисунок 5). СПА располагается на конце 3'-нетранслируемой области (НТО) длиной от ≈ 50 п.н. до нескольких тысяч пар нуклеотидов (т.п.н.) [40]. Каноническим СПА у млекопитающих является гексамер AAUAAA, встречающийся более чем в 60% мРНК человека. Вторым по частоте использования является СПА AUUAAA. Он представлен в 14% мРНК человека [63,64]. 8% мРНК содержат оба мотива AAUAAA и AUUAAA в последних 120 н. 15% мРНК человека используют в качестве СПА гексамеры, отличающиеся от AAUAAA или AUUAAA одним нуклеотидом [57]. 3.5% мРНК не содержат AAUAAA, AUUAAA и их однуклеотидных вариантов. Согласно данным МПРА СПА AAUAAA обеспечивает наиболее эффективный процессинг 3'-конца пре-мРНК. Гексамер AUUAAA в 2-3 раза менее эффективен. Все остальные варианты СПА обладают минимальной эффективностью [65]. Слабые СПА могут быть эффективны, если находятся в окружении других сильных цис-элементов (Рисунок 5) [69,70]. Мутации в СПА могут приводить к α - и β -талассемии, X-сцепленному синдрому иммунной дисрегуляции, полиэндокринопатии и энтеропатии (IPEX-синдрому), неонатальному диабету и предрасположенности к развитию рака [71–76]. Точечные мутации также могут приводить к появлению новых функциональных СПА, влияющих на уровень экспрессии гена [77].

У дрожжей СПА, также известный как элемент позиционирования, вырожден и представляет собой A-богатую последовательность [78]. Первые 2 аденина СПА у дрожжей распознаются по тому же механизму, что и у человека, консервативному для всех эукариот.

У растений мотив AAUAAA не столь консервативен, как у млекопитающих. Для них большее значение играет USE UUGUAW, напоминающий сайт связывания CFIm [79].

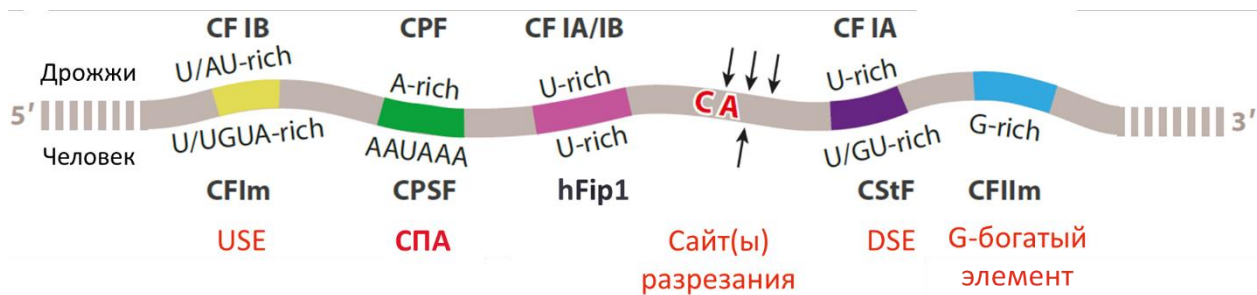


Рисунок 5. Схема участка пре-мРНК дрожжей и человека с цис-элементами, принимающими участие в процессинге 3'-конца по [68] с изменениями. CF – фактор разрезания; CPF – фактор разрезания и полиаденилирования; CPSF – фактор специфичности разрезания и полиаденилирования; CstF – фактор, стимулирующий разрезание; DSE – элемент последовательности, расположенный ниже СПА; USE – элемент последовательности, расположенный выше СПА. Чёрным цветом указаны белки, связывающие каждый элемент у дрожжей (сверху) и у человека (снизу).

Подготовка к сборке комплекса процессинга 3'-конца пре-мРНК начинается задолго до транскрипции РНК-полимеразой II функционального СПА. Комплекс mPSF (по крайней мере его субъединица CPSF160) рекрутируется к комплексу инициации транскрипции и сопровождает РНК-полимеразу II на протяжении всего процесса элонгации [80]. Таким образом, комплекс mPSF может принимать участие в координации всех стадий транскрипции [81].

На первоначальных этапах транскрипции РНК-полимераза II имеет наиболее высокий уровень фосфорилирования Ser5 (Ser5P) в CTD [54–56] (Рисунок 3). Ser5P, вероятно, мешает связыванию некоторых компонентов комплекса процессинга 3'-конца пре-мРНК [82]. Ser5P дефосфорилируется фосфатазой Ssu72 при переходе от инициации транскрипции к элонгации [57,83]. Для гистонов в теле гена характерно наличие метки H3K36me3, способствующей элонгации и терминции РНК-полимеразы II. Её добавляет метилтрансфераза SETD2. SETD2 также способствует фосфорилированию Ser2 в CTD [84]. Метка Ser2P накапливается по мере элонгации транскрипции и достигает своего пика в окрестности сайта

полиаденилирования [13] (Рисунок 3). Эту метку добавляют циклин-зависимые киназы 9 (CDK9) и 12 (CDK12) [59]. CDK9 также фосфорилирует фактор элонгации SPT5, способствуя эффективной элонгации [85], и 5'-3' экзорибонуклеазу XRN2, модулируя терминацию [59]. Накопление Ser2P в CTD способствует рекрутированию комплекса процессинга 3'-конца пре-мРНК и терминации [84]. У дрожжей фосфорилирование Tyr1 в CTD также накапливается по мере элонгации транскрипции, но резко снижается перед 3'-концом гена [13] (Рисунок 3). Tyr1P у дрожжей, предположительно, предотвращает преждевременное привлечение факторов терминации транскрипции на этапе элонгации [58]. У дрожжей Tyr1 и SPT5 дефосфолируются фосфатазой Glc7, способствующей терминации транскрипции [57,85–87]. У человека ортологом Glc7 является белковая фосфатаза 1 (PP1) [68]. Таким образом, ещё до транскрипции функционального СПА CTD РНК-полимеразы II имеет паттерн фосфорилирования, способствующий эффективному связыванию компонентов комплекса процессинга 3'-конца пре-мРНК и последующей терминации.

Комплекс CPSF

Сборка канонического комплекса процессинга 3'-конца пре-мРНК начинается с формирования гетеродимера CPSF160-WDR33 (Рисунок 4). Он является платформой для связывания остальных факторов терминации транскрипции [88]. После прочтения функционального СПА гексамер связывается комплексом из трёх субъединиц mPSF – CPSF160-WDR33-CPSF30. CPSF30 связывает нуклеотиды A1, A2, A4 и A5, а WDR33 – U3 и A6 в СПА [89]. CPSF160 не взаимодействует с пре-мРНК напрямую. Он выполняет роль скэффолда, рекрутируя WDR33 и CPSF30, и организуя их в правильную конфигурацию для связывания СПА. Комплекс CPSF160-WDR33-CPSF30 имеет высокую аффинность к молекулам РНК, содержащим канонический СПА AAUAAA [90,91]. Аффинность связывания РНК с

гексамером AUUAAA в 6 раз ниже по сравнению с каноническим СПА. Молекулы РНК, содержащие другие варианты СПА, имеют аффинность связывания минимум в 50 раз ниже, чем у гексамера AAUAAA [62]. CPSF30 связывает hFip1 – ещё одну субъединицу комплекса mPSF. hFip1 рекрутирует каноническую PAB, а также связывает U-богатые последовательности в области между СПА и сайтом разрезания пре-мРНК [92] (Рисунок 5). В клетках млекопитающих были идентифицированы 3 канонические PAB: PAB α и PAB γ /neo-PAB, экспрессирующиеся повсеместно [93,94], и PAB β , которая экспрессируется только в семенниках [95]. Канонические PAB нуждаются в каноническом комплексе процессинга 3'-конца пре-мРНК для связывания с РНК [96,97].

Комплекс mCF взаимодействует с mPSF, что приводит к формированию полноразмерного CPSF (Рисунок 4). В отличие от mCF, пространственная структура mPSF очень динамична [88]. В состав комплекса mCF входят 3 белка: CPSF73, CPSF100 и symplekin. Они взаимодействуют посредством своих CTD [98], формируя трёхлопастную структуру. Позиция каждой лопасти высоко вариабельна [88].

CPSF73 – эндонуклеаза, разрезающая пре-мРНК в сайте разрезания (полиаденилирования) [99,100]. Расстояние от СПА до сайта разрезания в большинстве случаев составляет 12-27 п.н. [62,101] Во многих источниках указано, что сайт разрезания располагается после динуклеотида CA [102,103] (Рисунок 5). Однако среди мРНК человека только 30% имеют динуклеотид CA перед сайтом разрезания. В 29% и 24% мРНК человека перед сайтом разрезания расположены динуклеотиды UA и GA соответственно. 17% мРНК человека не содержат аденин перед сайтом разрезания [62,104]. Согласно данным МПРА динуклеотид CA перед сайтом разрезания обеспечивает наиболее эффективный процессинг пре-мРНК у человека. Ему незначительно уступают по эффективности динуклеотиды UA и GA. Отсутствие аденина перед сайтом разрезания делает его малоэффективным [5]. Это обусловлено

пространственной структурой эндонуклеазы CPSF73. Аденин перед сайтом разрезания образует водородные связи с аминокислотами в составе активного центра CPSF73. Нуклеотид, расположенный перед аденином, не взаимодействует с CPSF73 [105]. Приблизительно у 1% европеоидов встречается однонуклеотидная замена, которая меняет последовательность сайта разрезания гена протромбина с CG на CA. Это приводит к значительному повышению эффективности разрезания пре-мРНК, повышению уровня экспрессии протромбина в 2 раза и, как следствие, склонности к тромбофилии [106,107]. Кристаллическая структура CPSF73 показывает, что в мономерной форме этот белок каталитически неактивен [108]. Для активации CPSF73 в составе комплекса mCF требуются обширные конформационные изменения всего комплекса [88,105]. Ингибиторы, связывающие активный центр CPSF73, имеют противораковый, противовоспалительный и антипротозойный эффекты [109–116].

Белок CPSF100 – паралог CPSF73 [100,117]. Он имеет мутацию в активном центре и, вероятно, лишён эндонуклеазной активности [108]. CPSF100 взаимодействует с CPSF160, обеспечивая сборку комплекса CPSF.

Symplekin имеет N-концевой домен (NTD) [118], с помощью которого он может взаимодействовать с белками Ssu72 [83,119,120], CstF64 [121] и CTD РНК-полимеразы II [62]. Ssu72 дефосфорилирует CTD РНК-полимеразы II в позиции Ser5 [83,122–124] и, в меньшей степени, Ser7 [57,120,125], способствуя связыванию других компонентов комплекса процессинга 3'-конца пре-мРНК [82]. Symplekin стимулирует фосфатазную активность Ssu72 [83,119].

Комплекс CstF

Комплекс CstF также, как и mCF, рекрутируется mPSF. Он состоит из трёх субъединиц: CstF64, CstF77 и CstF50 (Рисунок 4).

CstF64 содержит на N-конце РНК-связывающий домен типа RRM (мотив, распознающий РНК). RRM необходим и достаточен для распознавания DSE [65,126–128]. DSE располагается на расстоянии 10-30 п.н. ниже сайта разрезания [44] (Рисунок 5). У домена RRM CstF64 нет идентифицируемой консенсусной последовательности РНК, однако его связывание с РНК нельзя назвать неспецифичным. Роль DSE с разной эффективностью могут выполнять разнообразные последовательности [129]. CstF64 наиболее эффективно связывает GU- и U-богатые последовательности [127,129,130]. Домен RRM дополнен двумя α -спиралями на N-конце CstF64: N- и C-концевой. C-концевая спираль абсолютно консервативна у всех позвоночных. Она не принимает непосредственного участия в распознавании РНК. Напротив, она препятствует ему, закрывая собой поверхность RRM. При связывании с РНК C-концевая спираль разворачивается [126]. Предполагается, что последовательность CstF64 в ходе эволюции была оптимизирована не для образования сильных и уникальных контактов с определённой последовательностью, а для поддержания достаточной мобильности для связывания множества схожих последовательностей РНК [126].

Белок CstF64 имеет вторую изоформу – CstF64 τ . Белок CstF64 τ первоначально была обнаружена в мозге и семенниках [131]. Он принимает участие в модулировании альтернативного полиаденилирования при сперматогенезе [132]. Более поздние исследования показали, что CstF64 τ , вероятно, экспрессируется в гораздо большем количестве тканей и функционирует наряду с CstF64 [60,133]. CstF64 и CstF64 τ демонстрируют несколько различающуюся аффинность связывания различных РНК-полимеров. CstF64 имеет более высокую аффинность к U-богатым последовательностям, тогда как CstF64 τ с наибольшей эффективностью связывает GU-богатые последовательности [130]. CstF64 τ , в отличие от CstF64, не может взаимодействовать с symplekin [133]. Взаимодействие CstF64 с symplekin исключает взаимодействие с CstF77 и наоборот [121,134]. Мутации, нарушающие связывание CstF64 с symplekin, но при этом

сохраняющие его способность связываться с CstF77, не влияют на полиаденилирование, однако нарушают процессинг 3'-конца пре-мРНК гистонов [121]. Сила взаимодействия CstF64 с пре-мРНК влияет на частоту использования расположенного выше сайта полиаденилирования [135–137].

CstF77 взаимодействует с CstF64 и CstF50 [134,138–140], обеспечивая сборку всего комплекса. CstF77 способен к димеризации [138,141], что указывает на то, что комплекс CstF функционирует в форме димера. CstF77 взаимодействует с CPSF160 и WDR33, обеспечивая взаимодействие между комплексами CPSF и CstF [88].

Белок CstF50 так же, как и CstF77, способен к димеризации [134,142]. Роль CstF50 не вполне ясна. Предположительно, он может играть роль в ингибировании процессинга 3'-конца пре-мРНК при повреждении ДНК [143,144].

Связывая CPSF и DSE, CstF может влиять на выбор сайта разрезания и способствовать позиционированию пре-мРНК для последующего разрезания [68].

Комплекс CFII

Комплекс CFII_m может дополнительно увеличить специфичность комплекса процессинга 3'-конца пре-мРНК, связывая G-богатый элемент, расположенный ниже DSE некоторых пре-мРНК [145]. В состав CFII_m входят 2 белка: hPcf11 и hClp1 (полирибонуклеотидкиназа человека, субъединица 1 фактора разрезания) (Рисунок 4).

hPcf11 играет ключевую роль в сборке комплекса CFII_m, напрямую взаимодействуя с CstF, G-богатым элементом, CTD РНК-полимеразы II и hClp1 [42,146,147]. У дрожжей Pcf11 связывает как нефосфорилированный, так и фосфорилированный CTD, но имеет большее сродство к последнему [148–150]. hPcf11 взаимодействует с Ser2P в CTD в том случае, если Tyr1 дефосфорилирован [58,82,151]. Ser5P, вероятно, мешает связыванию Pcf11

[82]. У дрожжей взаимодействие Pcf11 с CTD не является необходимым для процессинга 3'-конца, но требуется для надлежащей терминации транскрипции [148,149].

Белок hClp1 обладает активностью 5'-ОН РНК-киназы [152]. Clp1 дрожжей способен связывать АТФ, однако не обладает ферментативной активностью [153–156]. Ферментативная активность hClp1 не является необходимой для активации эндонуклеазы CPSF73, но ряд свидетельств указывает на то, что связывание АТФ в активном центре Clp1 человека и дрожжей может быть необходимо для разрезания 3'-конца пре-мРНК [147,157–159].

Комплексы CstF и CFII_m необходимы для активации эндонуклеазы CPSF73, процессирующей 3'-конец пре-мРНК, однако механизм активации остаётся неизвестным.

Белок RBBP6

Предположительно, после связывания с пре-мРНК комплексов CPSF, CstF и CFII_m, в состав канонического комплекса процессинга 3'-конца пре-мРНК человека включается последний компонент, необходимый для активации его эндонуклеазной активности – белок RBBP6 (ретинобластома-связывающий белок 6) [157] (Рисунок 4). RBBP6 – неконститутивная субъединица комплекса CPSF [63,160]. По этой причине его роль в активации CPSF73 недооценивалась на протяжении многих лет. Убиквитин-подобный домен (UBL) белка RBBP6 и его дрожжевого ортолога Mre1 (экстрагенный супрессор мутантного Pcf11) взаимодействует с эндонуклеазой CPSF73 в составе комплекса процессинга 3'-конца пре-мРНК [157,160,161]. UBL Mre1 располагается вблизи активного центра эндонуклеазы Ysh1 (дрожжевой ортолог CPSF73) [161] и может направлять субстрат (молекулу пре-мРНК) в активный центр. Другой участок Mre1 взаимодействует с Pfs2 (субъединица фактора полиаденилирования, дрожжевой ортолог WDR33) и СПА [162]. Поэтому он был назван «участок,

чувствующий пре-мРНК» (PSR) Mpe1. Предполагается, что Mpe1 может чувствовать правильное связывание СПА другими субъединицами комплекса процессинга 3'-конца пре-мРНК с помощью своего PSR и передавать сигнал об узнавании РНК непосредственно эндонуклеазе Ysh1 с помощью домена UBL. Это, в свою очередь, может способствовать значительным структурным перестройкам комплекса процессинга 3'-конца пре-мРНК, необходимым для открытия и/или стабилизации активного сайта CPSF73 [105,157]. RBVP6, вероятно, схожим образом взаимодействует с mPSF [160]. Mpe1 также, предположительно, взаимодействует с PAP [162]. У человека RBVP6 после разрезания пре-мРНК остаётся связанным с 5'-продуктом разрезания [163]. Таким образом, RBVP6/Mpe1 может координировать реакции разрезания и полиаденилирования пре-мРНК.

В 2022 году впервые удалось воссоздать *in vitro* процессинг пре-мРНК с использованием рекомбинантных белков млекопитающих [157]. Было показано, что следующие компоненты являются необходимыми и достаточными для проведения реакции: комплексы CPSF, CstF, CFIm и белки RBVP6 и PAP, всего 14 белков. CTD РНК-полимеразы II не является обязательным. Также было показано, что комплекс CFIm стимулирует разрезание, но не является обязательным.

Комплекс CFIm

Комплекс CFIm состоит из двух копий белка CFIm25 и двух копий CFIm68 или CFIm59 (Рисунок 4). Его компоненты не имеют ортологов у *S. cerevisiae* и обнаружены только у многоклеточных эукариот [42]. CFIm связывается с РНК слабее остальных факторов процессинга 3'-конца пре-мРНК [157]. Предполагается, что этот комплекс принимает участие в процессинге далеко не всех молекул пре-мРНК. Основная функция CFIm – участие в регуляции альтернативного полиаденилирования [69,164,165]. Деpletion CFIm оказывает гораздо более сильное влияние на выбор сайтов полиаденилирования, чем depletion любого другого фактора процессинга

3'-конца пре-мРНК и приводит к увеличению доли транскриптов, процессированных в проксимальном сайте полиаденилирования [166–169]. Интересно отметить, что деплеция других факторов процессинга 3'-конца пре-мРНК приводит к увеличению доли транскриптов, процессированных в дистальном сайте полиаденилирования [170,171]. CFIm способствует выбору сайтов полиаденилирования, содержащих элемент USE с последовательностью UGUA, расположенный на расстоянии 10-40 п.н. выше СПА [165,172]. 56% мРНК человека содержат хотя бы один мотив UGUA в последних 120 н. [62]. CFIm связывает мотив UGUA в составе пре-мРНК и рекрутирует CPSF. В цитоплазме мотив UGUA может исполнять роль сайта связывания для белков Pumilio, тем самым способствуя деградации мРНК [173–178]. Таким образом, наличие мотива UGUA выше СПА может оказывать неоднозначное влияние на уровень зрелой мРНК. Оверэкспрессия компонентов CFIm может приводить к взаимодействию комплекса с субоптимальными сайтами, расположенными слишком близко к СПА. Это мешает связыванию CPSF с СПА и приводит к использованию расположенных ниже сайтов полиаденилирования [179,180].

CFIm25 связывает мотив UGUA [181]. CFIm59 и CFIm68 кодируются двумя генами-паралогами [182]. CFIm68 взаимодействует с hFip1, что обуславливает привлечение комплекса CPSF [183].

Сборка канонического комплекса процессинга 3'-конца пре-мРНК *in vivo* занимает около 10 сек и для «сильных» СПА происходит быстрее чем для «слабых» [184]. Сила СПА зависит от последовательности пре-мРНК и доступности коровых и вспомогательных факторов, способствующих или ингибирующих процессинг 3'-конца пре-мРНК [170].

Факторы, влияющие на эффективность сборки канонического комплекса процессинга 3'-конца пре-мРНК

Формирование вторичных структур РНК, включающих в себя цис-элементы, принимающие участие в процессинге 3'-конца, снижает

эффективность разрезания и полиаденилирования [185,186]. При этом, чем выше стабильность вторичной структуры, тем ниже эффективность процессинга 3'-конца. Это неудивительно, так как подавляющее большинство РНК-связывающих белков связывают оцРНК [187]. Однако образование вторичных структур РНК между СПА и сайтом разрезания, расположенным на расстоянии более 27 н., способствует процессингу 3'-конца [101]. Возможно, в данном случае вторичные структуры обеспечивают сближение СПА с сайтом разрезания до оптимального расстояния для сборки и функционирования комплекса процессинга 3'-конца пре-мРНК.

Сборка функционального комплекса процессинга 3'-конца пре-мРНК на СПА, расположенных в интронах, ингибируется факторами сплайсинга, мяРНК U1, SR-подобными ассоциированными с CTD факторами 4 и 8 (SCAF4, SCAF8), а также киназой CDK12 [188–191]. Наиболее хорошо изучена роль мяРНК U1 в подавлении активности интронных СПА, также известная как телескриптинг [190]. Сайты связывания мяРНК U1 aGGURAGu и GUGAGU могут препятствовать ранней терминации [192,193]. Телескриптинг не связан с участием мяРНК U1 в сплайсинге и обеспечивается высоким уровнем мяРНК U1, превышающем другие сплайсосомные мяРНК. Предположительно, это позволяет мяРНК U1 связать дополнительные сайты в интронах и супрессировать расположенные поблизости СПА. Сплайсинг терминального интрона, напротив, увеличивает эффективность расположенного ниже СПА [194–196]. Фактор U2AF, ассоциированный с 3'-сайтом сплайсинга, усиливает СПА посредством прямого взаимодействия с РАР [197]. U2-малый ядерный рибонуклеопротеин, ассоциированный с 3'-сайтом сплайсинга, помогает усилить процессинг 3'-конца пре-мРНК, взаимодействуя с CPSF [198]. Мутации в СПА могут ингибировать удаление терминального интрона [196]. Таким образом, терминация транскрипции и сплайсинг тесно связаны.

Полиаденилирование мРНК

После разрезания пре-мРНК РАР начинает безматрично добавлять к 3'-концу 5'-продукта разрезания аденины. Полиаденилирование мРНК разделено на две фазы. Во время первой фазы аденины добавляются с низкой эффективностью из-за слабого взаимодействия РАР с hFip1 [199]. РАР связывается с hFip1, добавляет несколько остатков аденозина и диссоциирует. Так продолжается до тех пор, пока поли(А)-хвост не достигнет минимальной длины, необходимой для связывания PABPN1 (ядерного поли(А)-связывающего белка). PABPN1 связывает около 10 поли(А) [200]. Вторая фаза полиаденилирования начинается с образования четырёхкомпонентного комплекса (РНК, РАР, CPSF, PABPN1), в котором CPSF и PABPN1 совместно способствуют переходу РАР в процессивное состояние, позволяющее синтезировать полноразмерный поли(А)-хвост без диссоциации [199,201]. Связывание PABPN1 увеличивает аффинность связывания РАР с РНК в 80 раз [202], при этом скорость добавления аденинов возрастает [203]. Предполагается, что PABPN1 ограничивает связывание CPSF с СПА и способствует взаимодействию между CPSF и РАР [204].

В процессе синтеза растущий поли(А)-хвост, связанный с PABPN1, сворачивается в сферу, способствуя контактам между CPSF, РАР и PABPN1. Когда поли(А)-хвост достигает длины в 250 н. у млекопитающих, новые молекулы PABPN1 больше не могут быть интегрированы в эту сферу и это приводит к разрушению взаимодействия между CPSF и РАР. В последствии РАР остаётся связанной с РНК только посредством взаимодействия с PABPN1, что приводит к возвращению в менее процессивное состояние и, в конечном счёте, остановке полиаденилирования [204]. Этот механизм обеспечивает регуляцию длины поли(А)-хвоста.

Длина образующегося в ядре поли(А)-хвоста является видоспецифическим признаком и составляет 200-250 н. у млекопитающих [205] и 70-90 н. у дрожжей [206,207]. Гипо- и гипераденилирование негативно

вливают на судьбу РНК. Гипоаденилированные транскрипты у млекопитающих остаются в ядре [208,209], у дрожжей они деградируют в ядре [210]. Деpletion PABPN1 в миобластах мыши приводит к укорочению поли(А)-хвостов и последующему накоплению транскриптов в ядре [209]. Полиаденилирование необходимо для созревания, стабильности и экспорта мРНК из ядра в цитоплазму [67]. Своевременный экспорт РНК из ядра в цитоплазму позволяет избежать её деградации в ядре [211,212].

Не так давно было открыто явление последовательного полиаденилирования мРНК в ядре [213–215], при котором пре-мРНК первоначально разрезается и полиаденилируется в дистальном сайте полиаденилирования. Затем та же молекула РНК разрезается и полиаденилируется комплексом процессинга 3'-конца пре-мРНК в проксимальном сайте полиаденилирования. На сегодняшний день точный механизм и распространённость последовательного полиаденилирования остаётся предметом исследований.

В цитоплазме PABPN1 замещается цитозольным поли(А)-связывающим белком (PABPC), который защищает мРНК от неспецифической деградации и способствует трансляции [216]. Одна копия PABPC покрывает ~30 н. поли(А) [217]. Средняя длина поли(А)-хвоста мРНК, расположенных в цитоплазме, составляет 30 н. у дрожжей и 50-100 н. у многоклеточных эукариот из-за действия деаденилаз Pan2-Pan3, CCR1-NOT и других [67,218]. Скорость деаденилирования и время полужизни транскриптов могут отличаться более чем в 1000 раз [219]. Время полужизни мРНК определяется её последовательностью и цитоплазматическими факторами. Мотивы в 3'-НТО объясняют 5.5% наблюдаемых вариаций во времени полужизни мРНК [220]. Поли(А)-хвост необходим для эффективной трансляции [221–224]. Эффективно транслирующиеся мРНК без поли(А)-хвоста, такие как мРНК гистонов, встречаются редко [225]. Поли(А)-хвост

может быть удлинён в цитоплазме [226,227]. Цитоплазматическое полиаденилирование регулирует трансляцию определённых мРНК [228].

Процессинг 3'-конца с помощью Star-PAP

Star-PAP (направляемая в спеклы PAP, регулируемая фосфатидилинозитол-4-фосфат-5-киназой типа 1 α) – ядерная неканоническая поли(А)-полимераза, кодируемая геном *TUT1* [229]. Она также обладает терминальной уридилтрансферазной активностью [230,231]. Star-PAP использует отличающийся механизм разрезания и полиаденилирования пре-мРНК (Рисунок 6).

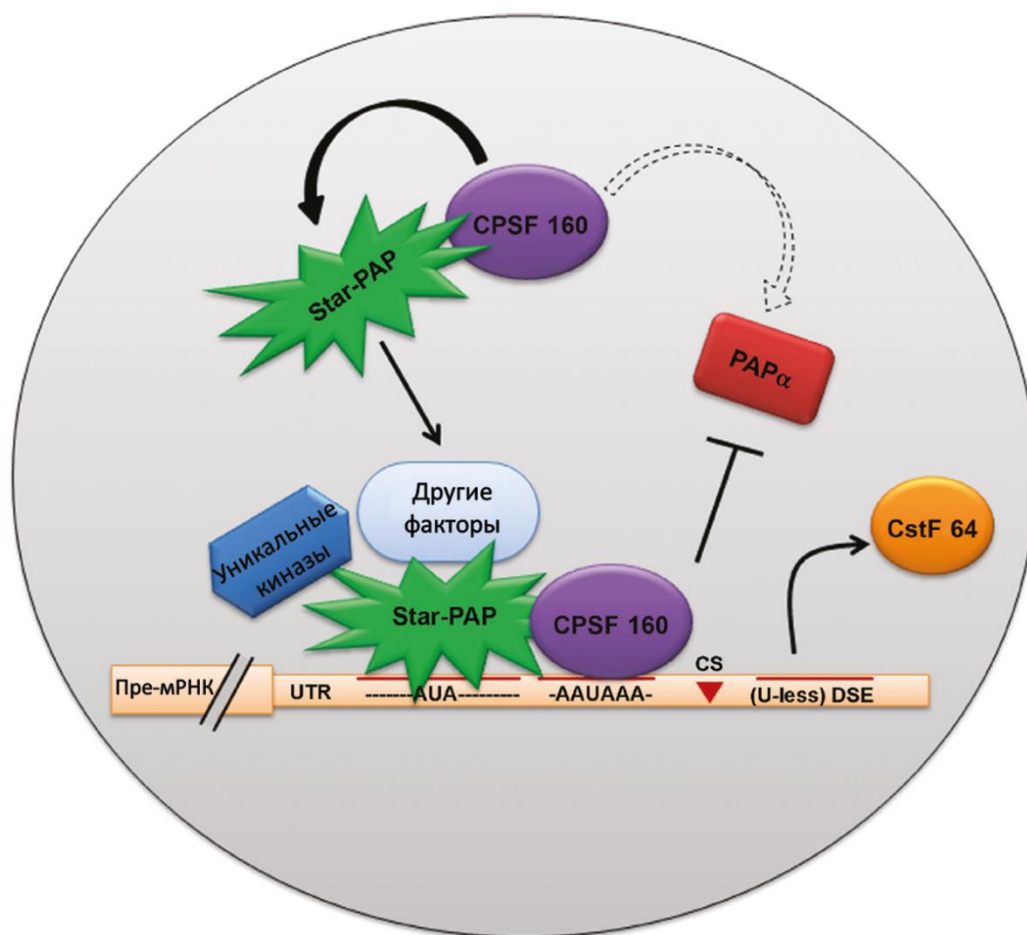


Рисунок 6. Модель выбора Star-PAP сайта полиаденилирования в соответствии с [232]. CPSF – фактор специфичности разрезания и полиаденилирования; CstF – фактор, стимулирующий разрезание; DSE – элемент последовательности, расположенный ниже СПА; PAP – поли(А)-полимераза; Star-PAP – направляемая в спеклы PAP, регулируемая фосфатидилинозитол-4-фосфат-5-киназой типа 1 α ; UTR – нетранслируемая область.

В отличие от канонических PAP, Star-PAP напрямую связывает пре-мРНК и рекрутирует комплекс CPSF для разрезания 3'-конца [233,234]. Star-PAP привлекается к пре-мРНК с последовательностью AUAyuu, расположенной в районе 50-150 н. выше СПА, и неоптимальной (U-бедной) последовательностью DSE [232]. По некоторым данным мотив AUA, окружён GC-богатыми последовательностями [234]. Star-PAP и PAP α конкурируют за связывание с CPSF-160, однако Star-PAP имеет преимущество [232]. Комплекс CstF не вовлечён в процессинг 3'-конца, опосредованный Star-PAP [232].

Было показано, что Star-PAP контролирует экспрессию большого количества мРНК, вовлечённых в разнообразные клеточные процессы, такие как ответ на окислительный стресс, апоптоз и онкогенез [229,233–240]. Даже при нокдауне Star-PAP PAP α не процессирует транскрипты Star-PAP и наоборот [229,233]. Однако известны случаи, когда Star-PAP и PAP α контролируют разные сайты полиаденилирования на одной молекуле пре-мРНК [235]. Если в гене, контролируемом Star-PAP, добавить в последовательность DSE Т-тракт длиной 6 п.н., процессинг его пре-мРНК будет контролироваться двумя PAP одновременно: PAP α и Star-PAP [232]. Последовательности 3'-НТО и DSE определяют, какой изоформой PAP будет контролироваться экспрессия того или иного гена [235].

Активность Star-PAP стимулируется индукцией двуцепочечных разрывов, окислительным стрессом и возрастает в 10 раз при связывании фосфатидилинозитол-4,5-бисфосфата [229,233,235]. Фосфатидилинозитол-4-фосфат-5-киназа типа 1 α (PIP5K1 α) и казеиновая киназа I α играет регуляторную роль в процессинге 3'-конца, опосредованном Star-PAP [229,234,241].

Альтернативное полиаденилирование

Альтернативное полиаденилирование (АПА) – процесс, при котором с одного гена образуются транскрипты, 3'-конец которых был процессирован с использованием различных СПА. Альтернативное разрезание – это процесс,

при котором транскрипты, использующие для процессинга 3'-конца один и тот же СПА, разрезаются и полиаденилируются в разных сайтах. У человека около 70% генов подвергаются альтернативному разрезанию и полиаденилированию [242]. В среднем у человека 5 сайтов полиаденилирования на ген [172] и 2,3 сайта разрезания на каждый сайт полиаденилирования [243].

АПА приводит к образованию мРНК с 3'-НТО разной длины, а в некоторых случаях к формированию различных изоформ белка с важными биологическими функциями. АПА имеет большое регуляторное значение. Оно влияет на включение в мРНК регуляторных районов, таких как сайты связывания микроРНК, регулирующие стабильность и трансляцию мРНК [244], а также AU-богатые элементы, способствующие стабилизации или деградации мРНК [245–249]. Примечательно, что выбор разных СПА может привести к привлечению в 3'-НТО белковых факторов, которые при трансляции могут напрямую связываться с образующимся белком и модулировать его функцию [250]. Также 3'-НТО могут содержать сигналы с определённой вторичной структурой, которые могут обуславливать специфическую цитоплазматическую локализацию в раннем развитии [251–253] или повышать стабильность транскрипта [225,254]. Тем не менее, не всегда ясно, какая доля событий АПА имеет физиологическую роль [170]. В культивируемых клетках человека НЕК293 ~10% всех случаев АПА подвержены посттранскрипционной регуляции и ~3% этих событий дестабилизированы микроРНК. У млекопитающих АПА в целом оказывает весьма слабое влияние на стабильность различных изоформ, уровень транскриптов [255–260] и эффективность трансляции [261]. У дрожжей, напротив, изменение длины 3'-НТО и даже замена одного нуклеотида может иметь большое влияние на стабильность транскрипта [254,262].

СПА, расположенные ближе к 5'-концу терминального интрона, имеют преимущество предположительно из-за того, что они первыми привлекают к

себе факторы процессинга 3'-конца, связанные с CTD РНК-полимеразы II [61]. Соотношение уровней проксимальных и дистальных транскриптов определяется относительной силой СПА и их окружения, а также скоростью деградации транскриптов [170].

Классический пример АПА – переключение между связанной с мембраной и секретируемой формой иммуноглобулина класса М (IgM) [263]. В ранних В-лимфоцитах снижен уровень белка CstF64. Это способствует выбору дистального СПА с более сильными последовательностями цис-элементов. При этом образуется мембранная изоформа IgM. По мере созревания В-лимфоцитов уровень CstF64 повышается, что способствует сборке канонического комплекса процессинга 3'-конца пре-мРНК на проксимальном СПА, окружённом более слабыми цис-элементами. При этом образуется укороченная секретируемая изоформа белка IgM [264,265].

Паузирование РНК-полимеразы II играет большую роль в выборе СПА. Паузирование РНК-полимеразы II между проксимальным и дистальным СПА способствует выбору проксимального СПА [266]. Увеличение или снижение скорости элонгации РНК-полимеразы II с помощью специфических мутаций приводит к глобальному удлинению или укорочению транскрипционных единиц РНК-полимеразы II соответственно [267]. Деплеция компонентов CFIm, PABPN1 и мРНК U1 снижает [268], а деплеция Pcf11, hFip1 и CstF64/CstF64 τ повышает [133,171] частоту использования дистальных СПА. PABPN1 способствует использованию дистальных сайтов полиаденилирования, конкурируя с CPSF за связывание со «слабыми» проксимальными гексамерами СПА [255,268]. При увеличении скорости пролиферации происходит сдвиг к проксимальным СПА у мышей и человека. При этом профили АПА отличаются в зависимости от типа клеток и стимула, используемого для увеличения скорости пролиферации. Механизм этого явления не ясен [269]. При дифференцировке миобластов C2C12 происходит сдвиг к дистальным СПА [270]. В индуцируемых плюрипотентных стволовых

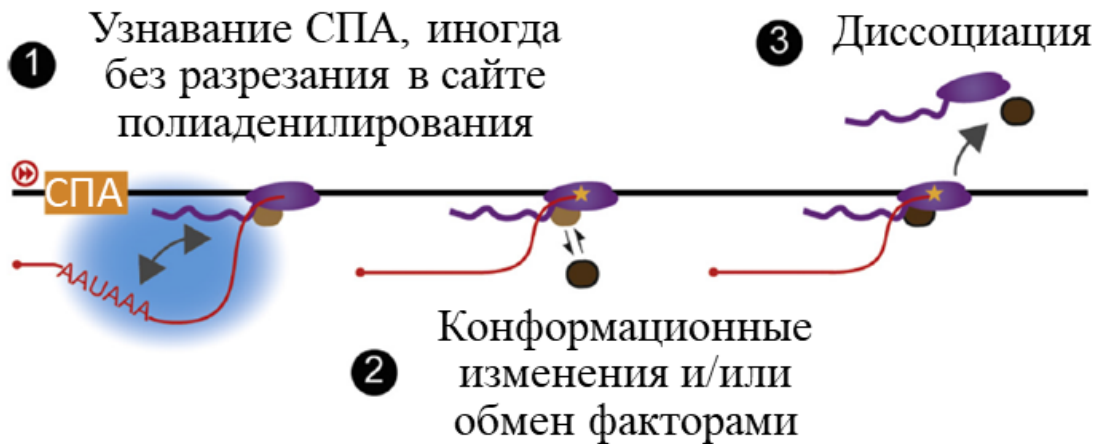
клетках происходит укорочение 3'-НТО [271]. На ранних этапах сперматогенеза происходит укорочение 3'-НТО генов, вовлечённых в созревание сперматозоидов и специфичных для семенников [272]. Есть примеры уникальных сайтов полиаденилирования, которые используются только в одной ткани, например в семенниках [273] или мозге [274]. В семенниках эти сайты часто имеют неканоническую последовательность гексамера или уникальные USE и DSE [273]. В мозге существуют тканеспецифичные сайты полиаденилирования с необыкновенно длинными 3'-НТО (>10 т.п.н.) [274]. Это потенциально предоставляет огромный репертуар регуляторных элементов, которые могут быть вовлечены в тканеспецифичную посттранскрипционную регуляторную сеть.

АПА главным образом тканеспецифично, нежели видоспецифично. Менее 1% сайтов полиаденилирования расположены в схожих участках генома человека, макаки резуса, мыши, крысы и собаки [242]. Профили АПА различаются у разных видов дрожжей [275]. В популяции гомогенных стволовых клеток также были обнаружены индивидуальные отличия в использовании сайтов полиаденилирования благодаря изучению отдельных клеток [276]. Это привело к заключению, что большая часть гетерогенности сайтов полиаденилирования связана с биологическим шумом (биологическими событиями, которые происходят *in vivo* как результат процесса с низкой специфичностью, который не может быть подвергнут эволюционной оптимизации) [275].

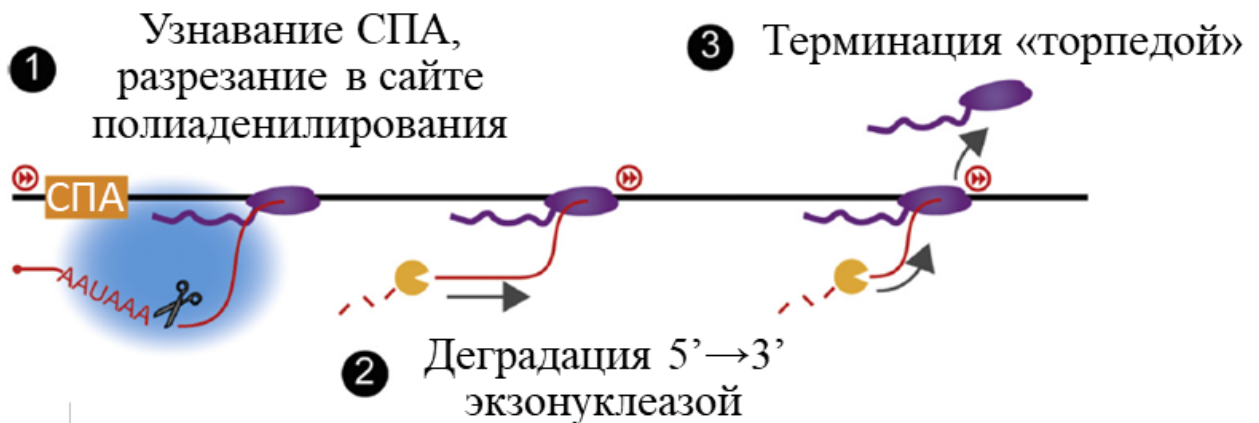
Терминация транскрипции РНК-полимеразы II в белок-кодирующих генах

На заре исследований терминации транскрипции были предложены две модели терминации транскрипции РНК-полимеразы II в белок-кодирующих генах: аллостерическая модель и модель торпеды (Рисунок 7) [277–279].

А Аллостерическая модель (с 1987)



Б Модель торпеды (с 1988)



В Унифицированная модель (с 2006)

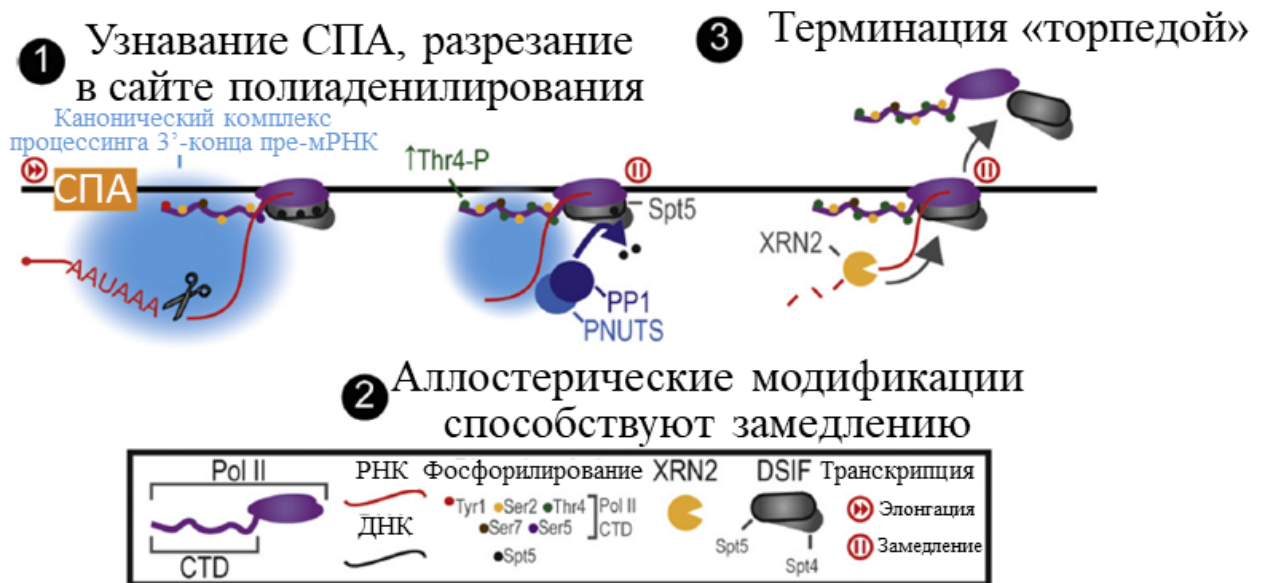


Рисунок 7. Модели терминации транскрипции РНК-полимеразы II в белок-кодирующих генах по [280]. А) Аллостерическая модель предполагает, что

изменение конформации РНК-полимеразы II и/или рекрутирование или диссоциация белковых факторов делают РНК-полимеразу способной к терминации.

Б) Модель торпеды предполагает, что разрезание пре-мРНК приводит к образованию незащищённого 5'-коца у молекулы РНК, ассоциированной с РНК-полимеразой II. Эта РНК деградируется в направлении $5' \rightarrow 3'$, что приводит к терминации.

В) Унифицированная модель предполагает, что разрезание пре-мРНК или сборка комплекса процессинга 3'-конца пре-мРНК способствует замедлению РНК-полимеразы II, которое обусловлено аллостерическим переходом, вызванным дефосфорилированием SPT5 комплексом PNUTS-PP1 и возросшим уровнем Thr4-Р в CTD. Этот аллостерический переход способствует паузированию РНК-полимеразы II и лёгкой терминации благодаря XRN2, деградирующей незащищённую с 5'-коца молекулу РНК, ассоциированную с РНК-полимеразой II после разрезания пре-мРНК. СПА – сигнал полиаденилирования; CTD – С-терминальный домен; DSIF – чувствительный к 5,6-дихлор-1- β -D-рибофуранозилбензимидазолу индуцирующий фактор; PNUTS – субъединица PP1, направляющая её в ядро; PP1 – белковая фосфатаза 1; Pol II – РНК-полимераза II; XRN2 – ядерная $5' \rightarrow 3'$ экзорибонуклеаза.

Аллостерическая модель предполагает, что элонгирующая РНК-полимераза II каким-то образом детектирует транскрипцию функционального СПА (Рисунок 7А). Это приводит к аллостерическим изменениям в самом ферменте и/или диссоциации неких факторов, способствующих элонгации или препятствующих терминации, или связыванию терминирующих факторов, таких как компоненты канонического комплекса процессинга 3'-конца пре-мРНК. Аллостерическая модель предполагает, что терминация транскрипции РНК-полимеразы II возможна без разрезания пре-мРНК.

Напротив, в модели торпеды разрезание пре-мРНК эндонуклеазой CPSF73 является ключевым событием, индуцирующим терминацию РНК-полимеразы II (Рисунок 7Б). При этом 3'-продукт разрезания пре-мРНК остаётся ассоциированным с элонгирующей РНК-полимеразой, однако его

5'-конец лишён кэпа и подвержен действию ядерной 5'→3' экзорибонуклеазы XRN2. XRN2 деградирует образующуюся РНК подобно тому, как торпеда «догоняет» РНК-полимеразу II, и вызывает терминацию транскрипции [281,282].

Обе эти модели не исключают друг друга. В последние годы растущую популярность набирает унифицированная модель, объединяющая аллостерическую модель и модель торпеды [283] (Рисунок 7 В). После прочтения СПА компонент комплекса DSIF фактор элонгации SPT5 дефосфорилируется комплексом PNUTS-PP1 [85,284,285]. PNUTS – субъединица PP1, направляющая её в ядро. Также после СПА возрастает уровень фосфорилирования Thr4 в CTD РНК-полимеразы II. Появление этой метки зависит от наличия функциональной эндонуклеазы CPSF73 [285–287]. Thr4P CTD может предвещать терминацию транскрипции или аллостерические изменения. После транскрипции функционального СПА скорость РНК-полимеразы II снижается с 2 т.п.н./мин до 100 п.н./мин и транскрипционный комплекс становится более склонным к терминации [284]. Это связано с ремоделированием комплекса элонгации и присутствием паузирующих последовательностей в ДНК [288,289]. У дрожжей после разрезания пре-мРНК РНК-полимераза II обычно транскрибирует не более 150-200 п.н. [196,290–292]. У млекопитающих расстояние от дистального СПА до сайта, где уровень транскрипции падает до фонового уровня – окно терминации – в среднем составляет 3,3 т.п.н. и может достигать 10 т.п.н. [293]. На окно терминации в среднем приходится 4 сайта терминации. Они обогащены мотивами $S_{2-6}AN_xW_{3-6}$ (Рисунок 8). Гибрид U-богатой РНК и A-богатой ДНК может вызывать обратный ход РНК-полимеразы II из-за низкой стабильности гетеродуплекса РНК:ДНК в активном центре фермента [14,15] (Рисунок 8А). Достигнув G/C-богатой последовательности РНК-полимераза II останавливается из-за высокой стабильности гибрида РНК:ДНК (Рисунок 8Б). Паузированная РНК-полимераза II является лёгкой мишенью для экзорибонуклеазы XRN2 [284].

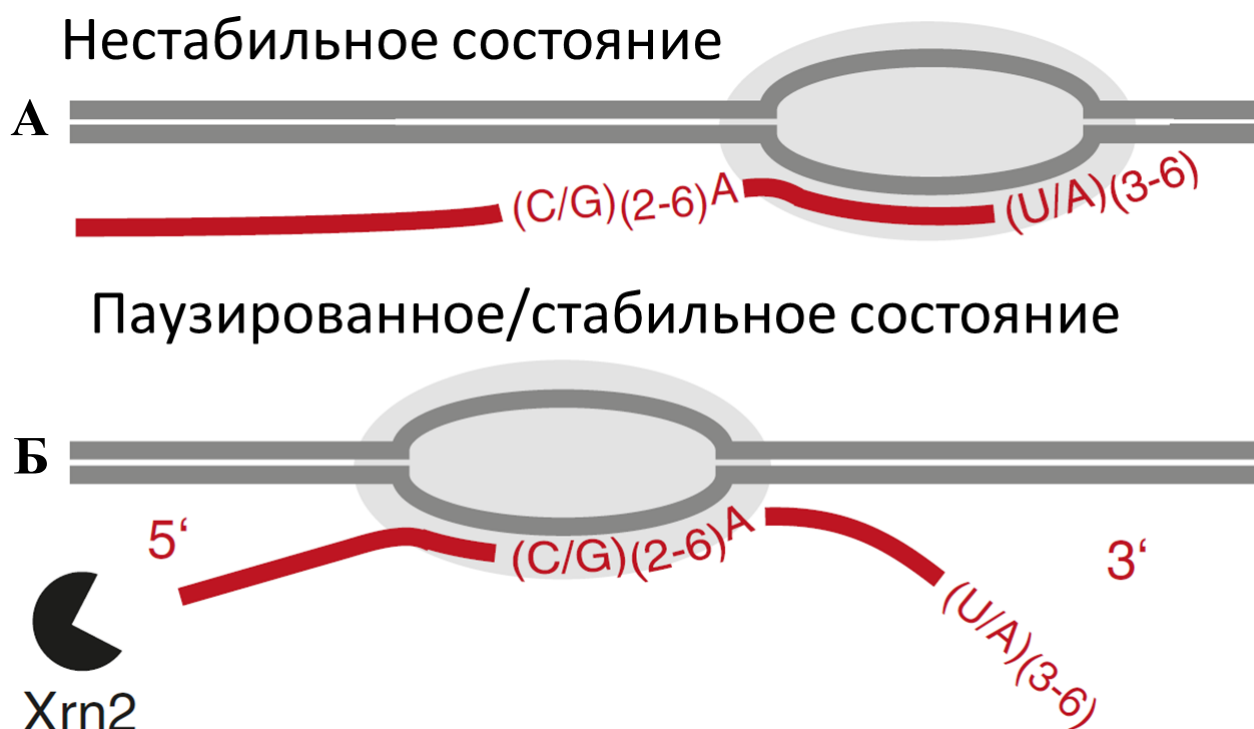


Рисунок 8. Сайт терминации транскрипции РНК-полимеразы II млекопитающих в соответствии с [293]. Предсказанные состояния РНК-полимеразы II в гибриде U-богатой РНК и А-богатой ДНК (**А**) и после обратного хода к сайту терминации транскрипции (**Б**). Xrn2 – ядерная 5'→3' экзорибонуклеаза.

Разрезание пре-мРНК, по-видимому, важно для высвобождения РНК-полимеразы II из тройного комплекса, так как мутации, приводящие к нарушению разрезания, также, как правило, приводят к нарушению терминации транскрипции [148]. После разрезания пре-мРНК ядерная 5'→3' экзорибонуклеаза XRN2 у человека или Rat1 у дрожжей деградирует лишённую кэпа образующуюся РНК [281,282]. XRN2 с лёгкостью «догоняет» замедленную РНК-полимеразу II, деградируя РНК вплоть до достижения активного центра фермента (Рисунок 8Б) [294]. Взаимодействие XRN2 с РНК-полимеразой II приводит к терминации транскрипции. У дрожжей Rat1 привлекается белком Rtt103, который взаимодействует с CTD РНК-полимеразы II, фосфорилированным по Ser2 и дефосфорилированным по Tyr1 [295]. РНК:ДНК геликаза Senataxin привлекается в R-петли и способствует терминации транскрипции. Предполагается, что, высвобождая РНК из гетеродуплекса с ДНК, Senataxin способствует активности XRN2

[296]. Таким образом, согласно унифицированной модели, терминацию транскрипции РНК-полимеразы II обеспечивают как аллостерические модификации фермента (дефосфорилирование SPT5), так и активность «торпеды» XRN2 (Рисунок 7В).

Однако в определённых условиях терминация транскрипции может происходить согласно аллостерической модели или модели торпеды (Рисунок 7А, Б). Эксперименты по терминации транскрипции *in vitro* показали, что наличия СПА достаточно для обеспечения терминации транскрипции РНК-полимеразы II даже в том случае, когда разрезание пре-мРНК не происходит [297]. При этом высвобождение РНК-полимеразы II из тройного комплекса происходит после конформационных изменений и является примером терминации транскрипции согласно аллостерической модели (Рисунок 7А). Тем не менее, терминация транскрипции РНК-полимеразы II при этом происходит на большем расстоянии от СПА, а образующаяся РНК лишена поли(А)-хвоста и деградирует в ядре. Терминацию транскрипции согласно модели торпеды (Рисунок 7Б) можно индуцировать добавлением антисмысловых олигонуклеотидов, вызывающих разрезание транскрипта РНКазой H1 [285,298,299]. Это приводит к деградации некэпированной образующейся РНК экзорибонуклеазой XRN2 и последующей терминации транскрипции. Антисмысловые олигонуклеотиды могут индуцировать терминацию транскрипции даже в отсутствии CPSF73 [285].

Нарушения терминации транскрипции у эукариот могут приводить к серьёзным последствиям для клетки [33]. Нетерминированная своевременно РНК-полимераза II может транскрибировать промотор расположенного ниже гена, что приведёт к снижению его активности. Это явление называется «транскрипционная интерференция» [300,301]. Если гены расположены «хвост-к-хвосту», нарушение терминации транскрипции может привести к образованию антисмысловых транскриптов, которые приведут к снижению уровня экспрессии расположенного ниже гена путём

РНК-интерференции [12]. Также может произойти столкновение между РНК-полимеразами, транскрибирующими разные цепи ДНК. При этом транскрипция обоих генов будет прервана [302,303]. Нетерминированная своевременно РНК-полимераза II может попасть в район репликации ДНК. Столкновение с ДНК-полимеразой может привести к нарушению синтеза ДНК и геномной нестабильности [304]. Особенно активно транскрибирующиеся гены имеют избыточные сигналы терминации на 3'-конце для предотвращения негативных последствий неэффективной терминации транскрипции [305–307]. РНК-полимераза II, избежавшая терминации, зависимой от канонического комплекса процессинга 3'-конца пре-мРНК, может быть терминирована альтернативными путями, которые, как правило, приводят к деградации транскрипта [308–311].

Альтернативные механизмы терминации транскрипции РНК-полимеразы II

Терминация транскрипции РНК-полимеразы II, зависящая от CoTC

Единственный известный на сегодняшний день альтернативный механизм терминации транскрипции РНК-полимеразы II, который приводит к образованию полиаденилированной РНК, – это терминация, зависящая от котранскрипционного разрезания (CoTC). Иногда РНК-полимераза II продолжает транскрипцию после прочтения функционального СПА на протяжении нескольких т.п.н. При прочтении АТ-богатой последовательности элемента CoTC происходит контранскрипционное разрезание пре-мРНК (Рисунок 9) [312]. При этом пре-мРНК остаётся связанной с элонгирующей РНК-полимеразой II посредством канонического комплекса процессинга 3'-конца пре-мРНК [313]. Затем происходит терминация транскрипции согласно модели торпеды, зависящая от XRN2 [282]. Комплекс РНК-полимеразы II с пре-мРНК высвобождается в нуклеоплазму, разрезается и полиаденилируется. CoTC может значительно увеличить выход зрелой мРНК и белка, так как обеспечивает процессинг 3'-конца пре-мРНК в нуклеоплазме вдали от аппарата деградации РНК, ассоциированного с хроматином [314].

CoTC нуждается в функциональном СПА [313]. CoTC впервые было обнаружено в генах β - и ϵ -глобина человека [313]. Предполагается, что CoTC играет роль в термине транскрипции 78 генов человека [315]. Точный механизм CoTC неизвестен.

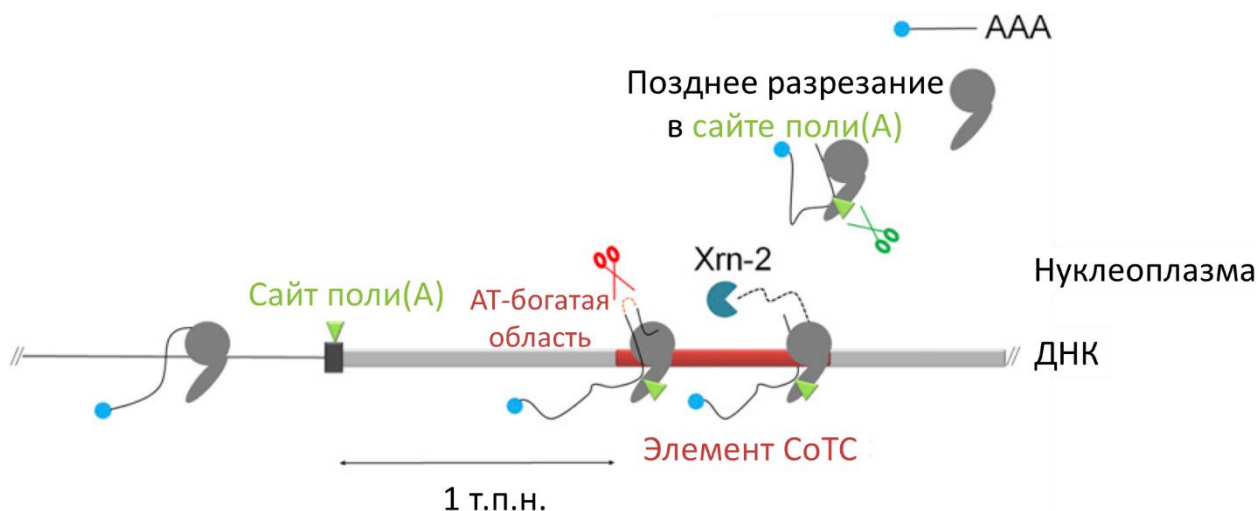


Рисунок 9. Терминация транскрипции РНК-полимеразы II, зависящая от CoTC в соответствии с [315]. При терминации транскрипции, зависимой от CoTC, транскрипты АТ-богатого элемента терминации (красный прямоугольник) разрезаются котранскрипционно (красные ножницы). Это способствует высвобождению РНК-полимеразы II до разрезания пре-мРНК в сайте полиаденилирования (зелёная стрелка) комплексом процессинга 3'-конца пре-мРНК (зелёные ножницы). CoTC – котранскрипционное разрезание; Xrn-2 – ядерная 5'→3' экзорибонуклеаза.

Терминация транскрипции в генах гистонов, зависящих от репликации

Гены гистонов, зависящих от репликации, не имеют интронов и кодируют короткие мРНК, которые не имеют поли(А)-хвоста. Эти гены расположены кластерами и транскрибируются в S-фазе клеточного цикла [316]. Специфический белковый комплекс, процессирующий 3'-конец пре-мРНК гистонов, распознаёт 2 цис-элемента – консервативную шпильку и расположенную на 15-20 н. ниже богатую пуринами последовательность (HDE) [317]. Шпилька и HDE обычно расположены в пределах 100 н. ниже стоп-кодона. Разрезание происходит между этими элементами, у

млекопитающих после пятого нуклеотида, расположенного ниже шпильки (как правило, это аденозин). Разрезание после других нуклеотидов происходит с меньшей эффективностью [318]. В результате образуется зрелая мРНК гистонов, на 3'-конце которой расположена шпилька и короткий одноцепочечный хвост. Шпилька и HDE распознаются белком, связывающим шпильку (SLBP) и малым ядерным рибонуклеопротеином (мяРНП) U7, соответственно [317,319]. Огромный белок, ассоциированный с FLICE (FLASH), взаимодействует с мяРНП U7 и комплексом разрезания пре-мРНК гистонов (HCC) [317,320,321]. В состав HCC входят белки CPSF73, CPSF100, CstF64 и symplekin. CPSF73 осуществляет роль эндонуклеазы, процессирующей 3'-конец пре-мРНК гистонов так же, как и в случае канонического процессинга 3'-конца пре-мРНК [105,316,322].

Терминация транскрипции комплексом Integrator (INT)

мяРНК являются ключевыми компонентами сплайсосомы и принимают участие в удалении интронов из пре-мРНК [323]. Гены мяРНК имеют на 3'-конце последовательность (3'-box РНК элемент), которую распознаёт комплекс INT. Он состоит, как минимум, из 14 субъединиц, включая белки с эндонуклеазной и фосфатазной активностью [324]. Субъединица комплекса INT (эндонуклеаза INTS11) разрезает РНК в области 3'-box РНК элемента [325,326]. Архитектура нуклеазного модуля INT схожа с mCF [324,327,328]. В частности, INTS11 имеет структурное сходство с CPSF73. Экзонуклеаза TOE1 процессирует 3'-концы мяРНК [329]. Последовательность 3'-box РНК элемента имеет консенсус GUUUUAAAAACAGAU [330]. Она гораздо более устойчива к заменам по сравнению с СПА. 3'-box РНК элемент активен только при транскрипции с промотора гена мяРНК [330,331].

Терминация транскрипции в генах гистонов и мяРНК не нуждается в активности XRN2 [267,332,333]. Вблизи 3'-конца генов гистонов, мяРНК и мякРНК происходит паузирование РНК-полимеразы II [334].

Терминация транскрипции комплексом NNS у дрожжей

У дрожжей терминацию транскрипции в генах мяРНК и мякРНК осуществляет комплекс NNS. Он состоит из 2 РНК-связывающих белков Nrd1 и Nab3 и геликазы Sen1 [335]. Nrd1 распознаёт последовательность GUAR, а Nab3 – расположенную ниже последовательность UCUUG [336–338]. Nrd1 также напрямую взаимодействует с Ser5P CTD РНК-полимеразы II [339–342]. Nrd1 и Nab3 обеспечивают позиционирование геликазы Sen1, которая взаимодействует с РНК и вытесняет РНК-полимеразу II подобно Rho-зависимой терминации у бактерий [343]. В результате образуется транскрипт, лишённый поли(А)-хвоста. После разрезания комплекс Nrd1-Nab3 рекрутирует экзосому, которая способствует быстрой деградации транскрипта. Деградация продолжается до тех пор, пока транскрипт не будет защищён РНК-связывающим белком (фактор сплайсинга для мяРНК или фактор модификации РНК для мякРНК). Таким образом экзосома способствует созреванию функциональных мяРНК [212,344,345]. Гены мякРНК за терминатором NNS несут терминатор канонического комплекса процессинга 3'-конца пре-мРНК [305–307]. Это защищает расположенные ниже гены от негативных последствий нарушений терминации транскрипции. Комплекс NNS также регулирует транскрипцию некоторых белок-кодирующих генов, способствуя преждевременной терминации транскрипции и, таким образом, снижая их уровень экспрессии [212,346–350]. У млекопитающих нет аналогов комплекса NNS. У дрожжей в транскрипцию и формирование 3'-конца мяРНК и мякРНК также вовлечён комплекс APT (Таблица 1) [351,352].

Проксимальная промоторная терминация

Большинство промоторов белок-кодирующих генов инициируют транскрипцию как в смысловом, так и в антисмысловом направлении [353–356]. В результате последней образуются транскрипты, известные как PROMPT (promoter upstream transcript). Транскрипция PROMPT обычно

терминируется в пределах 3 т.п.н. от сайта начала транскрипции. PROMPT быстро разрушаются ядерной экзосомой в направлении 3'→5' [353,357]. Похожие короткие и быстро деградируемые транскрипты также образуются в смысловом направлении белок-кодирующих генов в результате проксимальной промоторной терминации [357–359]. Комплекс INT вовлечён в проксимальную промоторную терминацию и терминацию антисмысловой транскрипции РНК-полимеразы II [360–363]. Также проксимальная промоторная терминация может обеспечиваться активностью комплекса Restrictor [364–367] и даже каноническим комплексом процессинга 3'-конца пре-мРНК при условии, что мРНК U1 не рекрутируется к пре-мРНК [190,191]. У дрожжей в проксимальную промоторную терминацию вовлечён комплекс NNS [339,341].

Терминация при «столкновении» РНК-полимеразы II с препятствием

У пекарских дрожжей транскрипционный фактор генов, кодирующих белки рибосом, Reb1 (белок, связывающий энхансеры РНК-полимеразы I) вызывает остановку РНК-полимеразы II [310]. Затем РНК-полимераза II убиквитинируется и, вероятно, деградируется протеасомой. РНК также деградирует с помощью экзосомы. Схожий механизм терминации транскрипции реализуется при столкновении РНК-полимеразы II с повреждениями ДНК [368].

Rnt1-зависимая терминация у дрожжей

У пекарских дрожжей ядерная рибонуклеаза двуцепочечной РНК Rnt1 (РНКазы III) распознаёт шпильку в РНК и разрезает её. 5'-продукт разрезания деградируется экзосомой [309]. 3'-продукт разрезания деградируется Rat1, что приводит к терминации согласно модели торпеды. Несколько генов дрожжей используют Rnt1 для СПА-независимой терминации транскрипции [308]. Два последних упомянутых механизма могут выполнять защитную функцию при «подтекании» терминации транскрипции, зависимой от канонического комплекса процессинга 3'-конца пре-мРНК [369].

Терминация, зависящая от Drosha

У пекарских дрожжей первичные микроРНК, полученные из длинных некодирующих РНК, котранскрипционно разрезаются белком Drosha. Разрезание способствует терминеции транскрипции, вероятно, зависящей от Rat1 [309,370].

Обратная торпеда

У дрожжей *Schizosaccharomyces pombe* РНК-полимераза II, избежавшая терминеции транскрипции, зависящей от канонического комплекса процессинга 3'-конца пре-мРНК, терминируется с помощью экзосомы [371]. При транскрипции комплекс элонгации может остановиться и отъехать назад, что делает 3'-конец РНК доступным для экзосомы. Далее экзосома, вероятно, деградирует РНК, что способствует обратному ходу РНК-полимеразы II и приводит к конформационным изменениям комплекса с последующей терминецией транскрипции. Такой механизм терминеции транскрипции был назван «обратная торпеда».

Основание J

Некоторые кинетопластиды конвертируют основание Т в ДНК в J. Основания J выполняют функцию терминаторов транскрипции РНК-полимеразы II. Вероятно, РНК-полимераза II при этом убиквитинируется и подвергается деградации [372].

Терминация в гене mod(mdg4) Drosophila melanogaster

У *D. melanogaster* и, по-видимому, других представителей семейства *Drosophilidae*, терминация транскрипции в четвёртом интроне гена *mod(mdg4)* обусловлена последовательностью длиной 79 н., формирующей специфическую вторичную структуру РНК [373]. Было показано, что не только вторичная структура, но и первичная последовательность этого района важна для обеспечения эффективной терминеции транскрипции. Предполагается, что терминация транскрипции происходит за счёт

распознавания этого района неким РНК-связывающим белком, специфичным для *Drosophilidae*. В результате образуется мРНК, лишённая поли(А)-хвоста [374].

Регуляция терминации транскрипции РНК-полимеразы II

Терминация транскрипции нарушается при стрессовом воздействии, вызванном повышением температуры, гипоксией, осмотическим шоком, окислительным стрессом, воздействием ультрафиолета и вирусными инфекциями [375–377]. Терминация транскрипции также может быть нарушена в процессе клеточного старения и при некоторых типах рака [378]. В большинстве случаев нарушения терминации транскрипции возникают в генах со слабыми СПА, расположенными в открытом хроматине вблизи других генов [59].

Заражение вирусом простого герпеса 1 типа (ВПГ-1) вызывает нарушение терминации транскрипции [377]. Это вызвано вирусным РНК-связывающим белком ICP27, который взаимодействует с CPSF и нарушает сборку функционального комплекса [379]. ICP27 также может связывать GC-богатые последовательности, расположенные выше СПА, удивительным образом рекрутируя CPSF и вызывая преждевременную терминацию транскрипции [380,381].

Вирус гриппа типа «А» также вызывает нарушение терминации транскрипции в заражённых клетках [382]. Неструктурный вирусный белок NS1A напрямую взаимодействует с CPSF30 [376,383,384]. Это не даёт комплексу CPSF связать свой РНК-субстрат [376]. NS1A также изолирует PABPN1, что приводит к укорочению длины поли(А)-хвоста до 12 н. в транскриптах хозяина [208]. Такие транскрипты остаются в ядре.

Нарушения терминации транскрипции также наблюдаются при некоторых типах рака и могут коррелировать с плохими прогнозами [378]. При светлоклеточном раке почки нарушения терминации транскрипции

коррелируют с мутациями в гене *SETD2*. Нарушения транскрипции могут приводить к возникновению новых химерных транскриптов и кольцевых изоформ [385], которые возникают из-за транс-сплайсинга и бэксплайсинга между соседними сонаправленными генами [386]. Формирование таких химерных транскриптов, вероятно, может приводить к образованию новых функциональных белков [59].

При тепловом шоке фактор теплового шока 1 (HSF1) напрямую взаимодействует с каноническим комплексом процессинга 3'-конца пре-мРНК [387] и может рекрутировать CPSF к генам, принимающим участие в ответе на тепловой стресс. Такие гены, как правило, короткие, имеют мало интронов и не подвержены телескриптингу [191].

У *Drosophila* наблюдается репрессия терминации транскрипции в кластерах РНК, взаимодействующей с Piwi (piРНК). Кластеры piРНК лишены независимых промоторов и транскрибируются за счёт нетерминированной транскрипции соседних генов [388–390]. Гены, окружающие кластер piРНК, расположены в ориентации «хвост к хвосту», что в случае нетерминированной транскрипции приводит к формированию двуцепочечной РНК. Поэтому кластеры piРНК несут метку H3K9me3. Эта метка привлекает в кластеры piРНК белок Rhino, ассоциированный с белком Cutoff. Rhino усиливает транскрипцию через кластеры piРНК, а Cutoff блокирует терминацию РНК-полимеразы II. Cutoff, предположительно, снижает активность XRN2 [267].

Терминация транскрипции РНК-полимеразы III

РНК-полимераза III отличается высокой процессивностью элонгации и очень эффективной терминацией. После терминации РНК-полимераза III способна быстро связываться с промотором того же гена для осуществления последующих раундов транскрипции (облегчённая реинициация) [391]. Традиционно считается, что РНК-полимераза III способна эффективно терминировать транскрипцию, достигнув Т-тракта переменной длины

(у дрожжей 6-7 нуклеотидов, у млекопитающих 4-5 нуклеотидов, часто меньше) [392,393]. Последовательности, фланкирующие Т-тракт, могут значительно влиять на эффективность терминации, особенно в терминаторах с короткими Т-трактами [394,395]. Однако консервативные последовательности, систематически встречающиеся ниже или выше Т-тракта, не были обнаружены. При транскрипции Т-тракта РНК-полимераза III останавливается. Предположительно, это связано с низкой термодинамической стабильностью ДНК:РНК гибрида в составе тройного комплекса [392]. В этом отношении механизм терминации транскрипции РНК-полимеразы III схож с Rho-независимой терминацией у бактерий и механизмом терминации транскрипции РНК-полимеразы I. Недавние исследования показали, что активный центр комплекса элонгации РНК-полимеразы III значительно слабее связывает РНК:ДНК дуплекс по сравнению с РНК-полимеразой II, что может объяснять особенно сильную чувствительность РНК-полимеразы III к слабому гибриду А-богатой ДНК с У-богатой РНК [14,15]. Недавние исследования указывают на то, что в терминацию транскрипции РНК-полимеразы III могут быть вовлечены белковые факторы [392] и расположенные выше У-тракта РНК-шпильки [396].

Использование МПРА и машинного обучения для анализа влияния последовательностей цис-элементов на экспрессию генов

МПРА позволяет исследовать влияние огромного количества (до нескольких миллионов) цис-элементов на уровень экспрессии, сплайсинг, полиаденилирование, трансляцию и распад РНК репортёрного гена [5] (Рисунок 10А). Для проведения МПРА создаётся библиотека, в которой интересующий участок репортёрного гена заменён на высоко вариабельную последовательность. Библиотеки доставляются в клеточные экстракты [397], клетки [398–404] или животных [405], в которых экспрессия репортёра приводит к возникновению молекулярного фенотипа (Рисунок 10Б).

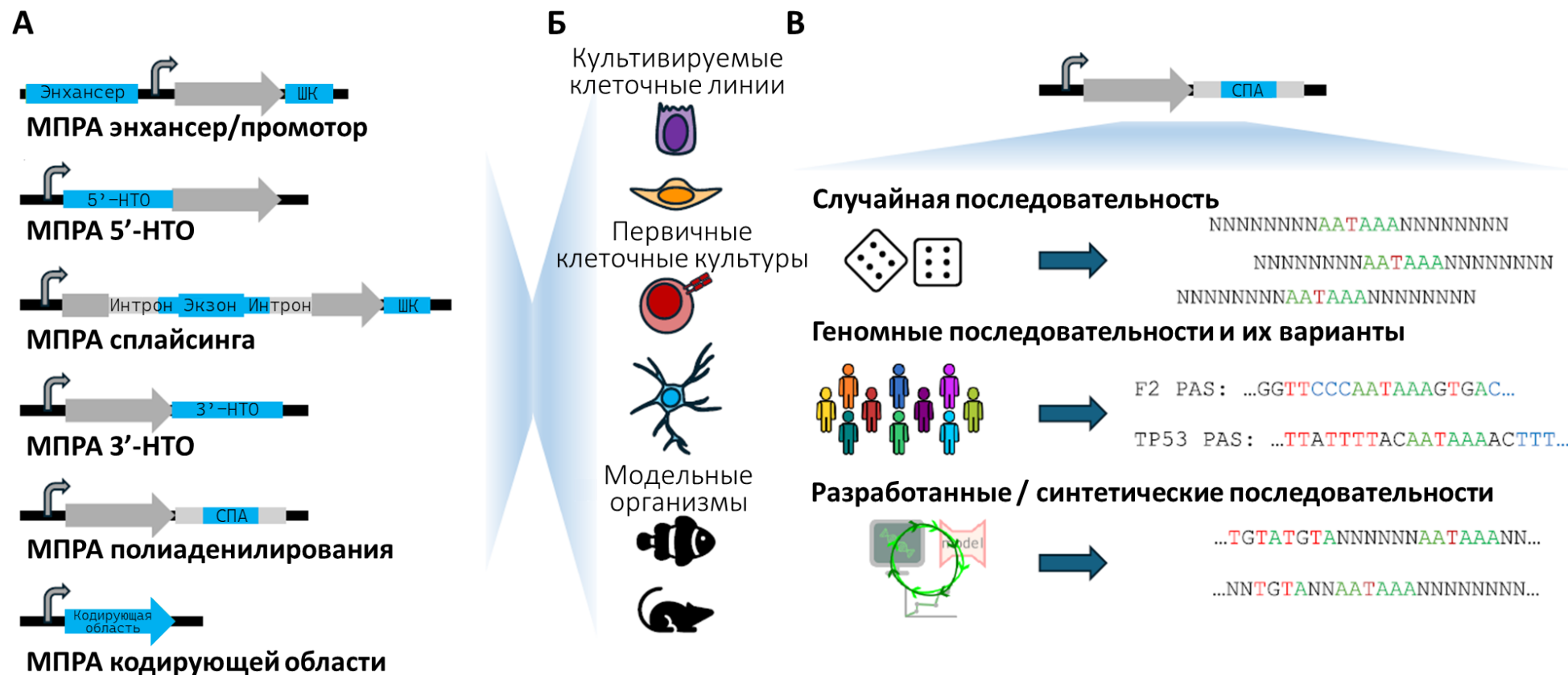


Рисунок 10. Форматы МПРА в соответствии с [406]. **А)** МПРА используют для анализа переменных последовательностей промоторов, энхансеров, экзонов, интронов, 5'-НТО и 3'-НТО. **Б)** МПРА проводят на клеточных экстрактах, культивируемых клеточных линиях, первичных клеточных культурах и модельных организмах, таких как бактерии, дрожжи и животные. **В)** Последовательности, исследуемые МПРА, могут быть случайными, типичными геномными последовательностями и их вариантами или разработанными вручную или с использованием разработанных алгоритмов. МПРА – массовый параллельный репортёрный анализ; НТО – нетранслируемая область; СПА – сигнал полиаденилирования; ШК - штрихкод.

Далее проводится количественный анализ репортёров и их молекулярных фенотипов с помощью высокопроизводительного секвенирования. МПРА использовали для анализа переменных последовательностей промоторов [397,398,400,407–409], энхансеров [402,405,410,411], экзонов [399,412], 5'-НТО [401,413–415] и 3'-НТО [5,6,403,416].

Данные МПРА очень хорошо подходят для обучения моделей с использованием методов машинного обучения. На практике модели и эксперименты развивались от исследований отдельных регуляторных элементов [398,402,405,408,413] до классического машинного обучения [401,417] и нейронных сетей [5,6,414,418–421].

Дизайн переменных последовательностей

Для проведения МПРА часто используются случайные нуклеотидные последовательности (Рисунок 10В). Это позволяет исследовать большое количество разнообразных последовательностей [412,413,417,422]. Конструирование библиотек случайных нуклеотидных последовательностей сравнительно дешево. В ходе таких экспериментов анализируют от нескольких сотен тысяч до миллионов переменных последовательностей [5,417,423]. Однако по мере увеличения длины переменной последовательности становится невозможным исследовать все её возможные варианты.

Для проведения МПРА также могут быть использованы геномные последовательности (Рисунок 10В) [403,409,410,416,424,425]. Библиотеки для МПРА такого типа могут содержать все возможные последовательности данного типа, аннотированные в геноме, и с высокой вероятностью обогащены последовательностями, имеющими «биологический смысл». Однако такие библиотеки обеднены последовательностями, которые снижают жизнеспособность. Схожий класс МПРА использует распространённые варианты геномных последовательностей [398,402,405,407]. В таких работах,

как правило, сравниваются молекулярные фенотипы типичной последовательности и её вариантов.

МППРА может быть проведён с использованием синтетических последовательностей, обогащённых специфическими цис-регуляторными мотивами или комбинациям мотивов для проверки гипотез о количестве, расстоянии и ориентации мотивов [6,400,426,427]. Также синтетические последовательности могут быть разработаны для выполнения определённой функции [5,414,420,428,429]. Последовательности могут быть разработаны вручную или с использованием моделей.

Анализ молекулярного фенотипа репортёров

Анализ молекулярного фенотипа репортёров может быть основан на активируемой флуоресценцией проточной сортировке клеток с последующим секвенированием ДНК [398,400]. При таком типе МППРА, переменные последовательности вносятся в ген флуоресцентного репортёра. Клетки сортируются по блокам на основе интенсивности флуоресценции, ДНК из каждого блока секвенируется, количество прочтений нормируется на общее количество прочтений в блоке [430,431] и активность каждой переменной последовательности определяется по распределению её нормированных прочтений по блокам. Репортёры часто интегрируют в геном клетки-хозяина в единственной копии, так как при таком подходе интенсивность флуоресценции клетки пропорциональна активности определённого репортёра [403,413,416,432]. Это особенно важно при работе с клетками млекопитающих, которые могут поглотить тысячи плазмид при временной трансфекции. Иногда для нормировки используется второй флуоресцентный белок. Такой подход к МППРА называют flow-seq [399], FACS-seq [413] или sort-seq [433]. Он позволяет определить влияние переменной последовательности ДНК на уровень репортёрного белка и может быть адаптирован для исследования различных этапов экспрессии гена от транскрипции до стабильности белка. Однако такой подход к МППРА зачастую

требует дополнительных исследований для определения того, на какой именно этап экспрессии гена влияет переменная последовательность. Схожий подход к МПРА в качестве молекулярного фенотипа использует скорость роста клеток в селективных условиях [434,435].

Ещё один распространённый класс МПРА использует секвенирование РНК для анализа транскриптов репортёрных генов [397,402,405,412]. При этом последовательности транскриптов не всегда содержат информацию о всех регуляторных элементах, представленных в ДНК репортёра. Поэтому для определения соответствия между транскриптом и репортёрным геном используют штрихкод – случайную или разработанную нуклеотидную последовательность. Соответствие между штрихкодом и исследуемой переменной последовательностью может быть определено секвенированием или задано в процессе создания библиотеки. Зачастую с каждой регуляторной последовательностью ассоциировано несколько штрихкодов для того, чтобы усреднить влияние штрихкода на регуляторную последовательность [436]. Это увеличивает размер библиотеки, а также стоимость её синтеза и секвенирования. Влияние регуляторного элемента определяется количеством прочтений соответствующего штрихкода, нормированным на представленность штрихкода в исходной библиотеке, количество прочтений альтернативной изоформы или количество прочтений в другой временной точке. Такой подход к МПРА совместим с временной трансфекцией и случайной интеграцией конструкций в геном. Он допускает использование библиотек РНК, транскрибированных *in vitro* [425].

Трансляционный МПРА объединяет в себе черты двух упомянутых выше типов МПРА [420,427,437]. Транскрипты репортёрных генов разделяют в градиенте сахарозы на фракции по числу связанных с ними рибосом и секвенируют. Уровень эффективности трансляции определяют по распределению количества прочтений последовательности по фракциям. Этот

подход может быть чувствителен к длине репортёра и наличию в кодирующей последовательности стоп-кодонов вне рамки считывания.

Анализ данных высокопроизводительного секвенирования

Данные МПРА нуждаются в корректной нормализации. Для этого могут быть использованы программы для анализа дифференциальной экспрессии, такие как DeSeq2 [438] и edgeR [439], или специализированные программы для анализа данных МПРА [440–442]. Зачастую исследователи разрабатывают собственные подходы для анализа экспериментальных данных МПРА.

После нормализации может быть проведён количественный анализ эффекта мотивов фиксированной длины [412,417,425] или расчёт позиционной весовой матрицы [411] для поиска потенциальных цис-элементов. На сегодняшний день достаточно популярным подходом является использование методов машинного обучения для анализа данных МПРА. Обученные модели могут предсказывать активности последовательностей, не проверенных экспериментально, использоваться для дизайна синтетических последовательностей и, дополненные методами интерпретации, давать информацию о взаимодействии различных цис-элементов.

Методы классического машинного обучения применяются для обучения моделей на данных МПРА, использующих в качестве входных данных характерные признаки последовательностей, такие, как представленность k-меров, включая линейную и логистическую регрессию [402,413,417,425,443–446], решающие деревья [447], градиентный бустинг [430,448] и случайный лес [449]. Модели классического машинного обучения отличаются сравнительной лёгкостью интерпретации.

В последние годы для обучения моделей на данных МПРА всё чаще используют методы глубокого обучения [5,6,414,418–420,450]. При этом одной из самых популярных архитектур являются свёрточные нейронные

сети. Увеличение глубины сети и использование обходных соединений может повысить производительность таких моделей [421,451]. Для обучения моделей на данных МПРА также могут быть использованы рекуррентные нейронные сети. Иногда модели комбинируют элементы свёрточных и рекуррентных нейронных сетей [452,453]. Ещё один класс моделей, используемых на данных МПРА, – большие языковые модели. Это модели, которых обучают без учителя реконструировать исходную белковую [454,455] или нуклеотидную [456–460] последовательность. После обучения слои большой языковой модели могут быть соединены с обучаемой с учителем нейронной сетью, предсказывающей активность переменных последовательностей, исследованных методом МПРА [461].

Дизайн последовательностей с использованием моделей, обученных на данных МПРА

Самый прямой путь для дизайна новых последовательностей с использованием моделей, обученных на данных МПРА – эволюция *in silico*. При этом в исходные последовательности шаг за шагом вносятся мутации. Последовательности с предсказанной активностью, наиболее близкой к целевой, сохраняются. Такой подход использовался для разработки 5'-НТО для генов дрожжей [414] и человека [420,449], а также энхансеров для культивируемых клеток *D. melanogaster* [462]. Основным недостатком этого метода является то, что он требует интенсивных вычислений.

Более эффективным подходом является использование метода градиентного спуска. Алгоритмы на его основе были использованы с моделями, обученными на данных МПРА, для создания различных регуляторных последовательностей [5,463,464].

Генеративно-состязательные сети и вариационные автокодировщики также были адаптированы для дизайна последовательностей [465–473]. При этом предварительно обученная на данных МПРА модель, предсказывающая активность последовательности, используется для обучения другой модели,

генерирующей последовательности с заданной активностью. Обучение генераторов последовательностей является вычислительно затратным, но они весьма эффективны. Для дизайна промоторов и энхансеров также могут быть использованы геномные авторегрессивные большие языковые модели [474]. Они удлиняют исходную нуклеотидную последовательность, выбирая нуклеотиды для удлинения на основе существующих геномных последовательностей, на которых была обучена большая языковая модель. Диффузионные модели также могут быть использованы для создания синтетических регуляторных элементов. В обучающие данные вносится шум. Диффузионные модели обучают удалять шум и восстанавливать исходную информацию обучающих данных [451,475–478]. Для создания новых последовательностей диффузионной модели на вход подают случайные последовательности, из которых она удаляет «шум», генерируя новые последовательности, которые схожи с последовательностями из обучающей выборки и с большой вероятностью играют регуляторную роль.

В конечном счёте, качество разработанных последовательностей зависит от качества модели, обученной на данных МПРА. Распространённым подходом для улучшения модели является увеличение объёма обучающей выборки. Например, последовательности могут быть сгенерированы на основе полученных ранее данных, исследованы экспериментально методом МПРА и использованы для улучшения модели. Такой подход позволил улучшить дизайн энхансеров, специфических для определённого типа клеток [479]. Для улучшения исходной модели также может быть использовано активное обучение. При активном обучении модель последовательно улучшается. Сначала она генерирует последовательности с неопределёнными предсказаниями, для них проводится МПРА, и полученные данные используются для обновления модели. Этот цикл повторяется несколько раз. Этот подход показал свою эффективность при обучении классификатора, предсказывающего активность энхансеров или сайленсеров последовательностей, содержащих сайт связывания специфического для

сетчатки глаза транскрипционного фактора колбочко-палочкового гомеобокса [480].

Исследование процессинга 3'-конца пре-мРНК с помощью МПРА и нейронных сетей

Цис-элементы, необходимые для разрезания и полиаденилирования пре-мРНК, расположены близко друг к другу и занимают достаточно короткий фрагмент РНК (Рисунок 5), что делает очень удобным исследование процессинга 3'-конца пре-мРНК с помощью МПРА.

Bogard, Linder и соавторы провели МПРА на культивируемых клетках человека HEK293 [5]. Библиотеки для МПРА содержали два последовательно расположенных сайта полиаденилирования генов человека или гена бычьего гормона роста (bGH). В большинстве библиотек участки последовательности длиной 20-25 п.н. вблизи проксимального и иногда дистального СПА были заменены на случайные нуклеотидные последовательности той же длины. Авторы секвенировали РНК репортёрного гена, определили соотношение проксимальных и дистальных продуктов полиаденилирования, а также сайты разрезания. Всего им удалось проанализировать более 3 миллионов последовательностей сайтов полиаденилирования. На этих данных авторы обучили свёрточную нейронную сеть APARENT (регрессионная сеть АПА). APARENT принимает на вход последовательности ДНК длиной 205 н., в которых СПА расположен в позиции 71..76 н. APARENT может предсказывать соотношение проксимальной и дистальной изоформ транскрипта. Для этого ему нужно предоставить две последовательности ДНК, а также указать расстояние между ними. APARENT также предсказывает сайты разрезания 3'-конца пре-мРНК и долю разрезанных по ним транскриптов. С помощью APARENT можно генерировать последовательности с заданным положением и эффективностью разрезания. При этом пользователь может до начала генерации зафиксировать последовательности определённых районов. APARENT в 90% случаев правильно предсказал влияние более 1000

однонуклеотидных мутаций генов человека из базы данных ClinVar [481], расположенных в области от 50 п.н. выше до 100 п.н. ниже аннотированного СПА.

Позже Linder с соавторами заново обработали данные того же эксперимента МПРА и обучили модель APARENT2 [421]. APARENT2 отличается от APARENT большим количеством слоёв и наличием обходных соединений. APARENT2 также принимает на вход последовательности ДНК длиной 205 н., в которых СПА расположен в позиции 71..76 н. Авторы адаптировали APARENT2 для анализа генов с одиночными СПА. APARENT2 превосходит производительность APARENT как на тестовых данных МПРА, так и на данных из базы GTEx о локусах количественных признаков, влияющих на выбор сайта полиаденилирования (3' ЛКП) [482,483]. Объединение APARENT2 с моделями, предсказывающими стабильность мРНК, дополнительно повысило её производительность.

Vainberg Slutskin с соавторами провели МПРА на культивируемых клетках миелогенной лейкемии человека K562 [6]. Библиотеки для МПРА содержали 12 тысяч варибельных последовательностей терминаторов транскрипции длиной 210 п.н., среди которых были нативные и мутированные последовательности терминаторов транскрипции человека и вирусов человека. На данных МПРА авторы обучили свёрточную нейронную сеть PolyApredictor. Она предсказывает уровень экспрессии гена на основе последовательности его сайта полиаденилирования, а также сайты разрезания 3'-конца пре-мРНК и их эффективность. Однако PolyApredictor демонстрирует низкую производительность по сравнению с APARENT2 и другими моделями [172,421].

Stroup с соавторами обучили модели PolyalD и PolyStrength на транскриптомных данных человека [172]. Модели принимают на вход последовательности длиной 240 н. Они имеют один свёрточный слой и один слой долгой краткосрочной памяти. PolyalD предсказывает вероятность того,

что подаваемая на вход последовательность является сайтом полиаденилирования. PolyStrength рассчитывает силу сайта полиаденилирования. Уникальной особенностью PolyID и PolyStrength является то, что они могут анализировать последовательности с нефиксированным положением СПА. Это позволяет использовать их для поиска потенциально функциональных СПА в геноме. Авторы обучили модель логистической регрессии, которая предсказывает, будет ли использован расположенный в интроне сайт полиаденилирования на основе предсказаний PolyStrength, расстояния до ближайших сайтов сплайсинга и их силе, размера окружающих экзонов и других признаков. Авторы также обучили модель логистической регрессии, которая предсказывает, будет ли использован расположенный в терминальном экзоне проксимальный сайт полиаденилирования на основе предсказаний PolyStrength для проксимального и дистального сайта полиаденилирования, расстояния между сайтами, размера расположенного выше интрона и расстояния до расположенного ниже гена. Авторы продемонстрировали, что PolyID и PolyStrength являются очень удобными инструментами для анализа сайтов полиаденилирования в геномном контексте. Однако PolyStrength не способен предсказать небольшие количественные изменения в силе сайта полиаденилирования. Авторы использовали PolyStrength для предсказания данных МПРА Bogard, Linder и соавторов [5]. PolyStrength смог эффективно классифицировать сайты полиаденилирования на сильные или слабые только в том случае, если частота использования дистального и проксимального сайта полиаденилирования отличалась более чем в 8 раз.

Wu и Bartel провели МПРА на культивируемых клетках человека НЕК293Т [101]. Они встроили между СПА и сайтом разрезания случайные нуклеотидные последовательности длиной от 1 до 37 п.н. Авторы проанализировали более 400 тысяч переменных последовательностей. Они показали, что оптимальное расстояние между СПА и сайтом разрезания составляет 15-18 н. В сайтах полиаденилирования, в которых расстояние

между СПА и сайтом разрезания составляло более 27 н., формирование вторичных структур между СПА и сайтом разрезания, по-видимому, способствовало более эффективному процессингу 3'-конца.

Kowalski, Wessels, Linder и другие разработали метод CPA-Perturb-seq [484]. Авторы создали библиотеку из 162 одиночных направляющих РНК, мишенью которых были 42 гена, регулирующих процессинг 3'-конца пре-мРНК. В их число вошли компоненты комплексов CPSF, CstF, CFIm, CFIIm, PAF (координирует инициацию и элонгацию транскрипции, контроль качества РНК) [485] и TREX (связывает транскрипцию РНК с ядерным экспортом) [486], а также фактор сплайсинга 3, богатый аргинином и серином (SRSF3). Авторы ингибировали экспрессию целевых генов в культивируемых клетках человека HEK293FT и K562. Затем они секвенировали РНК из одиночных клеток и проанализировали, как снижение уровня экспрессии генов, регулирующих процессинг 3'-конца пре-мРНК, влияет на уровень экспрессии и АПА мРНК генов человека. На полученных данных авторы обучили модель APARENT-Perturb. Она принимает на вход последовательности длиной 200 н. с расположенным посередине сайтом разрезания и предсказание APARENT2 для данного сайта полиаденилирования. APARENT-Perturb предсказывает частоту использования сайта полиаденилирования при снижении уровня экспрессии одного из 42 исследованных генов-регуляторов процессинга 3'-конца пре-мРНК. Этот эксперимент наглядно продемонстрировал взаимосвязь между определёнными цис-элементами и белками, принимающими участие в процессинге 3'-конца пре-мРНК.

Краткая история исследований DSE

Почти 50 лет назад юный Nicholas Proudfoot впервые смог синтезировать и секвенировать кДНК шести генов и обнаружил, что на расстоянии 15-21 н. выше поли(А)-хвоста располагается гексамер AAUAAA [1]. Вскоре благодаря развитию методов секвенирования и молекулярного клонирования стало

понятно, что, хотя наличие гексамера AAUAAA или его близких вариантов необходимо для обеспечения полиаденилирования 3'-конца пре-мРНК, на функциональность СПА также влияют окружающие его последовательности.

В частности, было показано, что GU-богатая последовательность, расположенная непосредственно после сайта полиаденилирования (DSE), способствует процессингу 3'-конца пре-мРНК [487–489]. Поначалу исследователи предполагали, что DSE по своей природе схож с СПА и существует одна или несколько консенсусных последовательностей DSE. В 80-х было опубликовано множество предполагаемых консенсусных последовательностей DSE: YGUGUUY [489], UGU [489,490], GUGUUG [491], CAYUG [492], UUAUUU [493–495], UUUUU [496], AGGUUUUU [497], UCCUGU [498]. Однако по мере увеличения числа секвенированных последовательностей стало понятно, что ни одна из предложенных консенсусных последовательностей, как и их возможные комбинации, не отражает наблюдаемое разнообразие DSE. Единственными наблюдаемыми общими чертами DSE были GU- и U-богатость [489,499].

Биохимические исследования показали, что DSE в пре-мРНК связывает комплекс CstF [65]. В 1997 году Beyer с соавторами очистили комплекс CstF из тимуса телёнка и культивируемых клеток человека HeLa, полученных из раковой опухоли шейки матки [129]. Они провели селекцию *in vitro* последовательностей РНК, которые связывает комплекс CstF. Оказалось, что в таких условиях комплекс CstF предпочтительно связывает РНК, содержащие одну из трёх консервативных последовательностей: AUGCGUCCUCGUCC, YGUGUYN₀₋₄UUYAYUGYGU, UUGYUN₀₋₄AUUUACUKN₀₋₂YCU. Эти мотивы встречались в некоторых известных DSE, однако по-прежнему не могли отразить всего разнообразия известных последовательностей. На тот момент уже было показано, что DSE может, по крайней мере частично, сохранять свою эффективность даже при наличии разнообразных мутаций [500–503]. Beyer и соавторы предположили, что сравнительно низкая

избирательность CstF по отношению к последовательности DSE *in vivo* обеспечивается за счёт белок-белковых взаимодействий между CstF и другими компонентами канонического комплекса процессинга 3'-конца пре-мРНК. При этом CstF связывает последовательности DSE с разной эффективностью, что может иметь регуляторную роль. Zarudnaya и соавторы предположили, что у DSE нет консенсусной последовательности и он может состоять из нескольких простых элементов длиной от 5 н., расположенных на разном расстоянии друг от друга [504]. DeKa и соавторы исследовали динамику РНК-связывающего домена CstF64 методом ядерного магнитного резонанса и предположили, что последовательность белка CstF64 была эволюционно оптимизирована не для того, чтобы формировать сильный контакт с уникальной последовательностью РНК, а, напротив, чтобы сохранять мобильность для связывания множества разнообразных, хотя и схожих последовательностей РНК [126].

21-й век ознаменовался началом геномной эры в биологии, что, вкупе с развитием биоинформатики, значительно расширило представление о разнообразии последовательностей, обеспечивающих процессинг 3'-конца пре-мРНК. В результате был обнаружен ряд новых цис-элементов с неизвестной функцией [165,243]. Hu с соавторами обнаружили, что DSE генов человека и мыши обогащены GU-, CU- и U-богатыми последовательностями [165], что было известно ранее. Они также показали, что GU-, CU-богатые мотивы представлены на расстоянии около 10 н. ниже сайта разрезания, тогда как U-богатые мотивы располагаются чуть дальше – на расстоянии около 25 н. ниже сайта разрезания. Позже Stroup и соавторы подтвердили, что в геноме человека проксимальный район DSE в большей степени обогащён GU, CU-богатыми мотивами, тогда как проксимальный – U-богатыми мотивами [172].

В 2011 году профессор Nicholas Proudfoot, совершивший множество открытий и ставший легендарной фигурой в области процессинга 3'-конца

РНК и терминации транскрипции, выступил с критикой существовавших на тот момент биоинформатических подходов к анализу геномных последовательностей, обеспечивающих процессинг 3'-конца пре-мРНК [61]. Он утверждал (и с этим трудно не согласиться), что многие принципы, лежащие в основе разрезания и полиаденилирования РНК, были известны задолго до начала геномной эры, а биоинформатические исследования в отрыве от экспериментальных подходов очень слабо расширили наши представления о механизмах процессинга 3'-конца пре-мРНК.

И тут на сцене появился МПРА...

Материалы и методы

Получение электрокомпетентных клеток *E. coli*

5 мл ночной культуры бактерий штамма TOP10 добавляли к 0.5 л питательной среды 2×YT (1.6% триптон, 1% дрожжевой экстракт, 0.5% NaCl). Клетки инкубировали на шейкере при 37°C до достижения оптической плотности ~0.7 при длине волны 600 нм. Затем инкубировали клетки на льду около 20 мин и центрифугировали при 4000 *g* 4°C 15 мин. Удаляли супернатант, ресуспендировали осадки в 1 л холодной ddH₂O и центрифугировали при 4000 *g* 4°C 15 мин. Удаляли супернатант, ресуспендировали осадки в 0.5 л холодной ddH₂O и снова центрифугировали при 4000 *g* 4°C 15 мин. Удаляли супернатант, ресуспендировали осадки в 20 мл 10% глицерина и центрифугировали при 4000 *g* 4°C 15 мин. Удаляли супернатант и ресуспендировали осадки в 2 мл 10% глицерола. Делали аликвоты клеток объемом 60 мкл. Хранили при -70°C.

Получение химически компетентных клеток *E. coli*

50 мл питательной среды Ψ_a (0.5% дрожжевой экстракт, 2% триптон, 0.5% MgSO₄), содержащей 1 мл ночной культуры бактерий штамма TOP10 или Trans1-T1 (TransGen), инкубировали на шейкере при 37°C до достижения оптической плотности 0.3 при длине волны 550 нм. Затем к 400 мл Ψ_a добавляли 20 мл полученной культуры бактерий и инкубировали на шейкере при 37°C до достижения оптической плотности 0.28 при длине волны 550 нм. Затем инкубировали клетки на льду 10 мин и центрифугировали при 500 *g* 4°C 15 мин. Удаляли супернатант и ресуспендировали осадки в 160 мл буфера TFB I (30 mM KOAc, 100 mM RbCl, 10 mM CaCl₂, 50 mM MnCl₂, 15% глицерин, pH = 5.8). Клетки охлаждали на льду в течение 5 мин и центрифугировали при 500 *g* 4°C 10 мин. Удаляли супернатант и ресуспендировали осадки в 16 мл буфера TFB II (10 mM MOPS, 75 mM CaCl₂, 10 mM RbCl, 15% глицерин,

pH = 6.5). Инкубировали клетки на льду 45-60 мин. Затем делали аликвоты полученных клеток объёмом 150 мкл. Клетки хранили при -70°C.

Трансформация плазмиды в электрокомпетентные клетки *E. coli*

Размораживали клетки на льду, добавляли 100 пг плазмиды, перемешивали пипетированием. Переносили клетки в кювету с шириной зазора между электродами 1 мм и проводили электропорацию при напряжении 1800 В и ёмкости 25 мкФ на приборе Gene Pulser Xcell (Bio-Rad). Добавляли 1 мл нагретой до 37°C питательной среды SOB (2% триптон, 0.5% дрожжевой экстракт, 8.56 mM NaCl, 2.5 mM KCl, 20 mM MgSO₄) без антибиотика и инкубировали на качалке при 37°C 1 час. Высевали на чашки с агаризованной средой (4% триптозный агар) и инкубировали при 37°C 16-18 часов.

Трансформация плазмиды в химически компетентные клетки *E. coli*

Размораживали клетки на льду, добавляли 100 пг плазмиды, перемешивали пипетированием и оставляли на льду на 30 мин. Затем инкубировали при 42°C 40 сек, после чего сразу переносили на лёд на 2 мин. Добавляли 1 мл нагретой до 37°C питательной среды питательной среды SOB (2% триптон, 0.5% дрожжевой экстракт, 8.56 mM NaCl, 2.5 mM KCl, 20 mM MgSO₄) без антибиотика и инкубировали при 37°C 1 час. Высевали на чашки с агаризованной средой (4% триптозный агар) и инкубировали при 37°C 16-18 час.

Выделение плазмидной ДНК (пДНК)

Плазмидные библиотеки выделяли с помощью набора GeneJET Plasmid Maxiprep (Thermo Fisher Scientific) в соответствии с рекомендациями производителя.

Одиночные плазмиды выделяли с помощью набора HiPure Plasmid Mini (Magen) или GeneJET Plasmid Midiprep Kit (Thermo Fisher Scientific) в соответствии с рекомендациями производителя.

Обработка пДНК экзонуклеазой T5

Чтобы избавиться от одноцепочечной кольцевой ДНК, 25 мкг плазмидной ДНК инкубировали в 1×буфере NEBuffer 4 (New England Biolabs) с 20 ед. экзонуклеазы T5 (New England Biolabs) в объёме 500 мкл при 37°C 16-18 часов. Останавливали реакцию добавлением 11.3 мкл 0.5M EDTA (pH = 8). Плазмиды очищали из реакционных смесей с помощью набора NucleoSpin Gel and PCR Clean-up (Macherey-Nagel) в соответствии с рекомендациями производителя и использовали для временной трансфекции клеток (см. далее).

Проведение ПЦР

Если продукт полимеразной цепной реакции (ПЦР) был предназначен для встройки в вектор, использовали полимеразу Phusion High-fidelity и буфер High-Fidelity (Phusion High-Fidelity PCR Kit, Thermo Fisher Scientific). Реакционная смесь объёмом 20 мкл содержала 100-400 пг пДНК, 1×буфер HF, 200-400 мкМ дНТФ, 200-500 нМ праймеры, 0.4 ед. ДНК-полимеразы Phusion High-fidelity. ПЦР проводили по следующей программе: 98°C 30-60 сек, (98°C 10-30 сек, 10-30 сек при температуре отжига праймеров, 72°C в течение 40 сек*длина продукта в т.п.н.)×35 циклов, 72°C 5 мин.

Для аналитической ПЦР мы использовали полимеразу HP-Taq (Биоспецифика). Реакционная смесь объёмом 20 мкл содержала около 200 пг плазмиды, 1×буфер (65 мМ трис-HCl, 16 мМ (NH₄)₂SO₄, 1.5 мМ MgCl₂, 0.05% Tween 20), 200 мкМ дНТФ, 1 мкМ праймеры, 1 ед. ДНК-полимеразы HP-Taq. ПЦР проводили по следующей программе: 96°C 5 мин, (96°C 30 сек, 30 сек при температуре отжига праймеров, 72°C в течение 1 мин*длина продукта в т.п.н.)×35 циклов, 72°C 5 мин.

Для аналитической ПЦР мы также использовали набор БиоМастер HS-Taq ПЦР-Color (Биолабмикс) в соответствии с рекомендациями производителя.

Объединение перекрывающихся фрагментов ДНК с помощью ПЦР (overlap extension PCR)

Для объединения перекрывающихся фрагментов ДНК с помощью ПЦР использовали полимеразу Phusion High-Fidelity и буфер High-Fidelity (Phusion High-Fidelity PCR Kit, Thermo Fisher Scientific). Реакционная смесь объёмом 19.2 мкл содержала 0.2 пМ каждого фрагмента ДНК, 1×буфер HF, 200 мкМ дНТФ, 0.4 ед. ДНК-полимеразы Phusion High-Fidelity. Первый раунд ПЦР проводили по следующей программе: 98°C 30 сек, (98°C 10 сек, 55°C 30 сек, 72°C в течение 40 сек*длина продукта в т.п.н.)×13 циклов, 72°C 5 мин. Затем в реакционную смесь добавляли праймеры до концентрации 200 нМ. Второй раунд ПЦР проводили по следующей программе: 98°C 30 сек, (98°C 10 сек, 30 сек при температуре отжига праймеров, 72°C в течение 40 сек*длина продукта в т.п.н.)×22 цикла, 72°C 5 мин.

Эмульсионная ПЦР (эПЦР)

Приготовление эмульсий и очистка продуктов проводились в соответствии с инструкцией к набору Micellula DNA Emulsion & Purification Kit (EUR_x). Смесь масла и ПАВ собирали, тщательно перемешивали и охлаждали до 4°C. Образцы водной фазы перемешивали на льду, как описано ниже. Эмульсионные ПЦР-продукты очищали на колонке после разрушения эмульсий путём добавления 1 мл 2-бутанола. ДНК элюировали в 50 мкл предварительно нагретого до 65°C элюирующего буфера из набора.

Однораундовая эПЦР

50 мкл смеси для эПЦР содержали 1 нг или 10 нг эквимольной смеси плазмиды №1 и плазмиды №2 в качестве матрицы, 1×буфер HF (Thermo Fisher Scientific), 200 мкМ дНТФ, 0.5 мкМ праймеров 20 и 29 (Приложение 2), 0.5 мг/мл бычьего сывороточного альбумина (БСА, EUR_x) и 2 ед. ДНК-полимеразы Phusion High-Fidelity (Thermo Fisher Scientific). Амплификацию проводили в следующих условиях: 95°C 1 мин, (95°C 10 сек, 55°C 30 сек, 72°C

10 сек)×25циклов, 72°C 5 мин. 10 мкл очищенных продуктов ПЦР анализировали с помощью электрофореза в 2% агарозном геле в буфере 0.5×TBE (45 mM трис-борат, 1 mM EDTA).

Двухраундовая эПЦР

Смеси для эПЦР раунда №1 объёмом 50 мкл содержали 10 пг, 100 пг или 1 нг эквимольной смеси плазмиды №1 и плазмиды №2 или 10 пг библиотеки 25-32 или 37-44 в качестве матрицы, 1×буфер HF (Thermo Fisher Scientific), 200 мкМ дНТФ, 0.5 мкМ одного из праймеров 20-28 и 0.5 мкМ праймера 29 (Приложение 2), 0.5 мг/мл БСА (EUR_x) и 2 ед. ДНК-полимеразы Phusion High-Fidelity (Thermo Fisher Scientific). Амплификацию проводили при следующих условиях: 95°C 1 мин, (95°C 10 сек, 55°C 30 сек, 72°C 10 или 30 сек)×15 циклов, 72°C 5 мин. 10 мкл очищенных продуктов ПЦР анализировали с помощью электрофореза в 2% агарозном геле в буфере 0.5×TBE. Смеси для раунда №2 эПЦР (50 мкл) содержали 0.3 или 0.5 мкл очищенных продуктов эПЦР раунда №1 в качестве матрицы, 1×буфер HF (Thermo Fisher Scientific), 200 мкМ дНТФ, 0.5 мкМ праймеров 49-50 (Приложение 2), 0.5 мг/мл БСА (EUR_x) и 2 ед. ДНК-полимеразы Phusion High-Fidelity (Thermo Fisher Scientific). Амплификацию проводили при следующих условиях: 95°C 1 мин, (95°C 10 сек, 52°C 30 сек, 72°C 10 или 30 сек)×18 или 20 циклов, 72°C 5 мин. 10 мкл очищенных продуктов ПЦР анализировали с помощью электрофореза в 2% агарозном геле в буфере 0.5×TBE.

Очистка продуктов ПЦР

Продукты ПЦР очищали с помощью наборов GeneJET PCR Purification (Thermo Fisher Scientific), PCR Purification (Qiagen), MinElute PCR Purification (Qiagen), HiPure PCR Pure Mini (Magen) или NucleoSpin Gel and PCR Clean-up (Macherey-Nagel) в соответствии с рекомендациями производителей.

Рестрикция пДНК и продуктов ПЦР

Для рестрикции использовали ферменты рестрикции FastDigest (Thermo Fisher Scientific), Time-Saver (New England Biolabs), Turbo (SibEnzyme) и соответствующие буферные растворы. Рестрикцию проводили в соответствии с рекомендациями производителей.

Электрофорез нуклеиновых кислот в агарозном геле

Электрофорез проводили в 0.7-2% агарозном геле в буфере 1×TAE (40 mM трис-ацетат, 1 mM EDTA) или 0.5×TBE с добавлением 0.5 мкг/мл бромистого этидия при напряжении 5В/см расстояния между электродами. Буфер для нанесения образцов содержал 30% глицерина, 0.25% бромфенолового синего и 0.25% ксиленцианола.

Выделение фрагментов ДНК из агарозных гелей

Для выделения ДНК из геля использовали наборы Gel Extraction (Thermo Fisher Scientific), HiPure Gel DNA mini (Magen) и NucleoSpin Gel and PCR Clean-up (Macherey-Nagel) в соответствии с рекомендациями производителей.

Отжиг комплементарных олигонуклеотидов

Смесь объемом 50 мкл содержала по 2 мкг каждого олигонуклеотида, 10 mM Tris-HCl (pH = 7.6), 50 mM NaCl и 1 mM EDTA. Инкубировали смесь при 95°C 5 мин, затем медленно остужали до комнатной температуры.

Лигирование плазмидного вектора с фрагментом ДНК

Для проведения реакции брали пДНК (50 нг) и ДНК встройки в мольном соотношении 1:3, 1 ед. ДНК-лигазы фага T4 (Биоспецифика) и 1×T4 DNA Ligase buffer (New England Biolabs). Доводили объем до 10 мкл ddH₂O. Инкубировали 16-18 часов при 4°C.

Одновременное лигирование более двух фрагментов ДНК

Для того чтобы одновременно лигировать более двух фрагментов ДНК, перекрывающихся между собой на 15-60 п.н., использовали Gibson Assembly Master Mix (New England Biolabs) согласно рекомендациям производителя.

Секвенирование ДНК по Сэнгеру

Реакционная смесь объемом 10 мкл содержала около 300 нг пДНК, 1×BigDye sequencing buffer (Thermo Fisher Scientific), 1 мМ праймер, 0.6 мкл BigDye Terminator v3.1 Ready Reaction Mix (Thermo Fisher Scientific). Реакцию проводили по следующей программе: 96°C 1 мин; (96°C 30 сек, 56°C 15 сек, 60°C 4 мин)×35 циклов.

По завершении реакции для очистки добавляли к реакционной смеси 15 мкл ddH₂O, 15 мкл 1М NaOAc и 60 мкл изопропанола. Инкубировали при комнатной температуре 20 мин, центрифугировали в течение 20 мин при 20000 g 4°C. Отбирали супернатант, осадок дважды промывали холодным 70% этанолом. Упаривали жидкость с помощью вакуумного концентратора Concentrator plus (Eppendorf).

Получение плазмидных библиотек для МПРА

Карты использованных в работе плазмид приведены в приложении 2. Для получения конструкции pTTC-hPGK-eGFP плазмиду pTTC-Hsap-WT [8] гидролизировали эндонуклеазами рестрикции FastDigest NheI и XbaI (Thermo Fisher Scientific) и лигировали саму на себя. Библиотеки для МПРА конструировали на основе плазмиды pTTC-hPGK-eGFP с помощью метода сборки Гибсона. Для каждой библиотеки соответствующие фрагменты ДНК амплифицировали с матрицы плазмиды pTTC-hPGK-eGFP с использованием праймеров 1-14 (Приложение 2). ПЦР проводили в объеме 50 мкл, содержащем 1 нг пДНК, 1×буфер HF (Thermo Fisher Scientific), 400 мкМ дНТФ, 500 нМ праймеров и 1 ед. ДНК-полимеразы Phusion High-fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C

1 мин, (98°C 30 сек, 60°C 30 сек, 72°C 3 мин)×35 циклов, 72°C 5 мин. Продукты ПЦР гидролизovali эндонуклеазой рестрикции DpnI (New England Biolabs) при 37°C в течение 4 часов в объёме 100 мкл для того, чтобы разрушить плазмидную матрицу и, таким образом, минимизировать примесь исходного нештрихкодированного вектора в библиотеках. После очистки с помощью набора PCR Purification (Qiagen), 76 фмоль (200 нг) вектора и 0.38 пмоль (~30 нг) встройки инкубировали с Gibson Assembly Master Mix (New England Biolabs) при 50°C 1 час. Затем реакционную смесь разбавляли в 10 раз ddH₂O и очищали MinElute PCR Purification (Qiagen). ДНК элюировали 10 мкл нагретого до 70°C буфера для элюции. Затем 5 мл очищенной реакционной смеси использовали для трансформации электрокомпетентных клеток *E. coli* TOP10. Плазмидные библиотеки для МПРА были получены н.с. лаборатории клеточного деления ИМКБ СО РАН Яринич Л.А.

Создание плазмидных конструкций с контрольными штрихкодами

Для создания контрольных конструкций, маркированных 20-буквенными штрихкодами (ШК), плазмиду pTTC-hPGK-eGFP гидролизovali эндонуклеазами рестрикции FastDigest XhoI и MluI (Thermo Fisher Scientific). Штрихкодированную последовательность дикого типа (WT) амплифицировали с матрицы плазмиды pTTC-hPGK-eGFP с использованием праймеров 15 (ШК I) или 16 (ШК II) и 19 (Приложение 2). Штрихкодированную последовательность с делецией цитозина в позиции +32 п.н. после СПА (ΔC) амплифицировали с плазмиды pTTC-hPGK-eGFP-ΔC с использованием праймеров 17 (ШК III) или 18 (ШК IV) и 19 (Приложение 2). ПЦР проводили в объёме 50 мкл, содержащем 1 нг пДНК, 1×буфер HF (Thermo Fisher Scientific), 400 мкМ дНТФ, 500 нМ праймеров и 1 ед. ДНК-полимеразы Phusion High-fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C 1 мин, (98°C 30 сек, 57°C 30 сек, 72°C 30 сек)×35 циклов, 72°C 5 мин. Продукты ПЦР гидролизovali эндонуклеазами рестрикции FastDigest XhoI и MluI (Thermo Fisher Scientific) при 37°C в

течение 1 часа. После очистки с помощью набора PCR Purification (Qiagen), 50 нг вектора и 7 нг встройки лигировали при 4°C в течение ночи Т4 ДНК-лигазой (Евроген). Затем 1 мкл реакционной смеси использовали для трансформации электрокомпетентных клеток *E. coli* TOP10.

Создание нештрихкодированных плазмидных конструкций с индивидуальными мутациями

Для создания нештрихкодированных плазмидных конструкций с индивидуальными мутациями в плазмиду рТТС-Hsap-WT [8] ввели точечную мутацию в сайт эндонуклеазы рестрикции SpeI, расположенный в промоторе фосфоглицераткиназы 1 человека (hPGK), обеспечивающего транскрипцию репортёра eGFP. Для этого плазмидный вектор рТТС-Hsap-WT гидролизировали эндонуклеазами рестрикции FastDigest SpeI и EcoRI (Thermo Fisher Scientific). С плазмиды рТТС-Hsap-WT амплифицировали фрагмент промотора hPGK и кодирующую последовательность eGFP с использованием праймеров 19 и 53 (Приложение 2). ПЦР проводили в объёме 50 мкл, содержащем 1 нг пДНК, 1×буфер HF (Thermo Fisher Scientific), 400 мкМ дНТФ, 500 нМ праймеров и 1 ед. ДНК-полимеразы Phusion High-fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C 1 мин, (98°C 30 сек, 58°C 30 сек, 72°C 50 сек)×35 циклов, 72°C 5 мин. Продукты ПЦР гидролизировали эндонуклеазами рестрикции FastDigest XbaI и EcoRI (Thermo Fisher Scientific) при 37°C в течение 1 часа. После очистки фрагментов из агарозного геля с помощью набора Gel Extraction (Thermo Fisher Scientific), 50 нг вектора и 45 нг встройки лигировали при 4°C в течение ночи ДНК-лигазой фага Т4 (Биоспецифика). 1 мкл реакционной смеси использовали для трансформации электрокомпетентных клеток *E. coli* TOP10.

Методом ОТ-кПЦР (обратной транскрипции - количественной ПЦР) было показано, что введённая мутация не влияет на активность промотора. Модифицированную плазмиду рТТС-Hsap-WT гидролизировали по уникальным сайтам эндонуклеазы рестрикции FastDigest SpeI и BsiWI (Thermo Fisher

Scientific) и чистили с помощью набора PCR Purification (Qiagen). Для каждой мутации, два одноцепочечных комплементарных нуклеотида отождели друг на друга (Приложение 2, праймеры 54-75). В результате образовались двуцепочечные фрагменты ДНК с липкими концами сайтов рестрикции SpeI и BsiWI. 50 нг вектора и 0.3 нг встройки лигировали при 4°C в течение ночи T4 ДНК-лигазой (Евроген). Затем 1 мкл реакционной смеси использовали для трансформации электрокомпетентных клеток *E. coli* TOP10 или весь объём лигазной смеси для трансформации химических компетентных клеток *E. coli* Trans1-T1 (TransGen).

Для создания нештрихкодированных плазмидных конструкций с индивидуальными мутациями и репортёром люциферазой (NanoLuc), плазмидные конструкции с последовательностью WT или одной из мутаций GTGTACTT и TCAGATAC гидролизовали эндонуклеазами рестрикции FastDigest SmaI и EcoRI (Thermo Fisher Scientific) при 37°C в течение 1 часа. Кодировующую последовательность люциферазы амплифицировали с плазмидной матрицы (которая была любезно предоставлена сотрудниками лаборатории иммуногенетики ИМКБ СО РАН) с использованием праймеров 89-90 (Приложение 2). ПЦР проводили в объёме 50 мкл, содержащем 1 нг пДНК, 1×буфер HF (Thermo Fisher Scientific), 400 мкМ дНТФ, 500 нМ праймеров и 1 ед. ДНК-полимеразы Phusion High-fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C 1 мин, (98°C 30 сек, 58°C 30 сек, 72°C 30 сек)×35 циклов, 72°C 5 мин. Продукты ПЦР гидролизовали эндонуклеазами рестрикции FastDigest SmaI и EcoRI (Thermo Fisher Scientific) при 37°C в течение 1 часа. После очистки с помощью набора HiPure PCR Pure Mini (Magen), 50 нг вектора и ~8 нг встройки лигировали при 4°C в течение ночи T4 ДНК-лигазой (Евроген). 1 мкл реакционной смеси использовали для трансформации электрокомпетентных клеток *E. coli* TOP10.

Для создания нештрихкодированных плазмидных конструкций с терминаторами SV40 early и bGH модифицированную плазмиду

pTTC-Hsap-WT гидролизовали эндонуклеазами рестрикции FastDigest EcoRI и MluI (Thermo Fisher Scientific) при 37°C в течение 1 часа и чистили из геля с помощью набора HiPure Gel DNA mini (Magen). Последовательность терминатора SV40 early амплифицировали с матрицы плазмиды pTTC-Hsap-WT с помощью праймеров 76-77 (Приложение 2). Последовательность терминатора bGH амплифицировали с матрицы плазмиды pSpCas9(BB)-2A-GFP (Addgene #48138) с помощью праймеров 82-83 (Приложение 2). ПЦР проводили в объёме 50 мкл, содержащем 1 нг пДНК, 1×буфер HF (Thermo Fisher Scientific), 400 мкМ дНТФ, 500 нМ праймеров и 1 ед. ДНК-полимеразы Phusion High-fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C 1 мин, (98°C 30 сек, 58°C 30 сек, 72°C 30 сек)×35 циклов, 72°C 5 мин. Продукты ПЦР гидролизовали эндонуклеазами рестрикции FastDigest EcoRI и MluI (Thermo Fisher Scientific) при 37°C в течение 1 часа. После очистки с помощью набора PCR Purification (Qiagen), 50 нг вектора и ~18 нг встройки лигировали при 4°C в течение ночи T4 ДНК-лигазой (Биоспецифика). 1 мкл реакционной смеси использовали для трансформации электрокомпетентных клеток *E. coli* TOP10. Нештрихкодированные плазмидные конструкции с терминаторами SV40 early и bGH были получены с.н.с. лаборатории клеточного деления ИМКБ СО РАН Андреевой Е.Н.

Для создания нештрихкодированных плазмидных конструкций с точечными мутациями терминаторов SV40 early и bGH модифицированную плазмиду pTTC-Hsap-WT гидролизовали эндонуклеазами рестрикции PacI (Thermo Fisher Scientific) 16 часов и Turbo EcoRI (SibEnzyme) 3 часа при 37°C и чистили из геля с помощью набора HiPure Gel DNA mini (Magen). Фрагменты последовательностей терминаторов амплифицировали с помощью праймеров 78-81, 84-87 (Приложение 2). ПЦР проводили в объёме 20 мкл, содержащем 100 пг пДНК, 1×буфер HF (Thermo Fisher Scientific), 200 мкМ дНТФ, 500 нМ праймеров и 0.4 ед. ДНК-полимеразы Phusion High-fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C 30 сек, (98°C

10 сек, 62°C 30 сек, 72°C 5 сек)×25 циклов, 72°C 5 мин. Перекрывающиеся на 24-25 п.н. фрагменты последовательностей терминаторов объединяли между собой с помощью ПЦР. Во втором раунде ПЦР использовали праймеры 88-89 (Приложение 2). Продукты ПЦР гидролизовали эндонуклеазами рестрикции PacI (Thermo Fisher Scientific) 16 часов и Turbo EcoRI (SibEnzyme) 3 часа при 37°C. После очистки с помощью набора NucleoSpin Gel and PCR Clean-up (Macherey-Nagel), 50 нг вектора и ~14 нг встройки лигировали при 4°C в течение ночи T4 ДНК-лигазой (Биоспецифика). Весь объем реакционной смеси использовали для трансформации химических компетентных клеток *E. coli Trans1-T1* (TransGen).

Все плазмидные конструкции проверяли секвенированием по Сэнгеру.

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 CO₂ 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 $0.5-1 \times 10^6$ культивируемых клеток в лунку шестилуночного планшета (9.6 см²). 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.

Флуоресцентная микроскопия

Трансфицированные клетки были проанализированы непосредственно перед сборкой для выделения тотальной РНК. Все образцы были

сфотографированы при одних и тех же настройках с использованием микроскопа Axio Vert.A1 (Zeiss), оснащённого камерой AxioCam ICm1 (Zeiss).

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 в 1×буфере для ДНКазы I с добавлением $MgCl_2$ (Thermo Fisher Scientific) с 3 ед. ДНКазы I (Thermo Fisher Scientific) и 40 ед. эндонуклеазы рестрикции DpnI (New England Biolabs) при 37°C 30 мин чтобы избавиться от примесей гДНК и пДНК. Затем очищали РНК с помощью набора CleanRNA Standard (Евроген) согласно рекомендациям производителя. Реакцию обратной транскрипции проводили в объёме 20 мкл, содержащем 1-3 мкг очищенной тотальной РНК, 500 мкМ дНТФ, 2.5 мкМ праймера олиго(dT)₂₀ или 2 пМ праймера PacI-R (Приложение 2), 1×буфер First-Strand или RT (Thermo Fisher Scientific), 40 ед. RNaseOut Recombinant RNase Inhibitor (Thermo Fisher Scientific) и 100 ед. обратной транскриптазы RevertAid (Thermo Fisher Scientific). Контрольный образец без обратной транскриптазы (no reverse transcriptase control, NRT) содержал все указанные выше компоненты, кроме обратной транскриптазы RevertAid и RNaseOut Recombinant RNase Inhibitor. Реакционную смесь инкубировали при 42°C 1 час, затем при 70°C 10 мин.

Подготовка образцов МПРА к секвенированию

Все плазмидные библиотеки разводили до концентрации 100 нг/мкл и добавляли к ним плазмидные конструкции с контрольными штрихкодами до концентрации 0.25 нг/мкл. Подготовка всех образцов МПРА включала в себя 2 раунда ПЦР. Во время первого раунда каждый образец МПРА метили уникальным восьмибуквенным индексом, входящим в состав одного из праймеров 20-28, 31-48 (Приложение 2, индексы выделены красным). Индексы были разработаны н.с. лаборатории клеточного деления ИМКБ СО РАН Иванкиным А.В. Во время второго раунда ПЦР к образцам присоединяли

адаптеры P5 и P7 для секвенирования на платформе Illumina MiSeq, входящие в состав праймеров 49-50 (Приложение 2, выделены красным).

Для подготовки образцов картирования в качестве матрицы использовали плазмидные библиотеки, разведённые до 10 пг/мкл. Первый раунд ПЦР проводили в объёме 50 мкл, содержащем 10 пг пДНК, 1×буфер HF (Thermo Fisher Scientific), 200 мкМ дНТФ, 250 нМ одного из праймеров 20-28, 250 нМ праймера 29 (Приложение 2) и 1 ед. ДНК-полимеразы Phusion High-Fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C 30 сек, (98°C 10 сек, 55.8°C 30 сек, 72°C 30 сек)×5 циклов, (98°C 10 сек, 72°C 30 сек)×10 циклов, 72°C 5 мин. Продукты ПЦР очищали с помощью набора GeneJET PCR Purification (Thermo Fisher Scientific) согласно рекомендациям производителя и элюировали дважды 76 мкл буфера для элюции, предварительно нагретого до 70°C. Смешивали по 5 мкл очищенных продуктов I раунда ПЦР из первой и второй элюции и разбавляли в 24 раза ddH₂O. Второй раунд ПЦР проводили в объёме 50 мкл, содержащем 4 мкл очищенного разбавленного продукта I раунда ПЦР, 1×буфер HF (Thermo Fisher Scientific), 200 мкМ дНТФ, 250 нМ праймеров 49-50 (Приложение 2) и 1 ед. ДНК-полимеразы Phusion High-fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C 30 сек, (98°C 10 сек, 60.9°C 30 сек, 72°C 30 сек)×5 циклов, (98°C 10 сек, 72°C 30 сек)×13 циклов, 72°C 5 мин. Продукты ПЦР очищали с помощью набора GeneJET PCR Purification (Thermo Fisher Scientific) согласно рекомендациям производителя.

Для подготовки образцов нормирования в качестве матрицы использовали плазмидные библиотеки, разведённые до 2.5 нг/мкл. Первый раунд ПЦР проводили в объёме 50 мкл, содержащем 2.5 нг пДНК, 1×буфер HF (Thermo Fisher Scientific), 200 мкМ дНТФ, 250 нМ праймера 30, 250 нМ одного из праймеров 31-48 (Приложение 2) и 1 ед. ДНК-полимеразы Phusion High-Fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C 30 сек, (98°C 10 сек, 70.5°C 10 сек, 72°C 5 сек)×5 циклов, (98°C

10 сек, 72°C 10 сек)×10 циклов, 72°C 5 мин. Второй раунд ПЦР проводили в объёме 100 мкл, содержащем 2 мкл реакционной смеси I раунда ПЦР, 1×буфер HF (Thermo Fisher Scientific), 200 мкМ дНТФ, 250 нМ праймеров 49-50 (Приложение 2) и 2 ед. ДНК-полимеразы Phusion High-fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C 30 сек, (98°C 10 сек, 60.9°C 10 сек, 72°C 6 сек)×4 цикла, (98°C 10 сек, 72°C 12 сек)×6 циклов, 72°C 5 мин. Продукты ПЦР очищали с помощью набора GeneJET PCR Purification (Thermo Fisher Scientific) согласно рекомендациям производителя.

Для подготовки образцов экспрессии в качестве матрицы использовали кДНК. Первый раунд ПЦР проводили в объёме 50 мкл, содержащем 4 мкл кДНК (1/5 объёма реакционной смеси для проведения ОТ), 1×буфер HF (Thermo Fisher Scientific), 200 мкМ дНТФ, 250 нМ праймера 30, 250 нМ одного из праймеров 31-48 (Приложение 2) и 4 ед. ДНК-полимеразы Phusion High-fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C 30 сек, (98°C 10 сек, 70.5°C 10 сек, 72°C 5 сек)×5 циклов, (98°C 10 сек, 72°C 10 сек)×10 циклов, 72°C 5 мин. Второй раунд ПЦР проводили в объёме 200 мкл, содержащем 4 мкл реакционной смеси I раунда ПЦР, 1×буфер HF (Thermo Fisher Scientific), 200 мкМ дНТФ, 250 нМ праймеров 49-50 (Приложение 2) и 4 ед. ДНК-полимеразы Phusion High-fidelity Hot Start (Thermo Fisher Scientific). ПЦР проводили по следующей программе: 98°C 30 сек, (98°C 10 сек, 60.9°C 10 сек, 72°C 6 сек)×4 цикла, (98°C 10 сек, 72°C 12 сек)×6 циклов, 72°C 5 мин. Продукты ПЦР очищали с помощью набора GeneJET PCR Purification (Thermo Fisher Scientific) согласно рекомендациям производителя.

Для точного измерения концентрации ДНК в образцах МПРА для высокопроизводительного секвенирования использовали количественную ПЦР в реальном времени (кПЦР) (см. ниже).

Проведение кПЦР

кПЦР проводили с использованием набора БиоМастер HS-qPCR SYBR Blue (Биолабмикс), амплификатора CFX96 Touch Real-Time PCR Detection System (Bio-Rad) и праймеров 51, 52, 92-97 (Приложение 2). Реакционная смесь объемом 20-25 мкл содержала 250-400 нМ праймеры. кПЦР проводили по следующей программе: 95°C 5 мин, (95°C 15-30 сек, 60°C 30-45 сек, 72°C 0-30 сек)×40 циклов. Детекцию проводили на каждом раунде амплификации.

Для точного измерения концентрации ДНК в образцах МПРА для высокопроизводительного секвенирования использовали стандарты Illumina DNA (КАРА) и праймеры 51, 52 (Приложение 2). Для измерения уровня экспрессии eGFP в клетках, трансфицированных одиночными плазмидами, использовали праймеры 95 и 96 (Приложение 2). Для измерения уровня экспрессии mCherry в клетках, трансфицированных одиночными плазмидами, использовали праймеры 97 и 98 (Приложение 2). Уровень экспрессии eGFP нормировали на уровень экспрессии mCherry. При каждом запуске в качестве отрицательных контролей использовали образцы без матрицы (no template control, NTC) и без обратной транскриптазы (NRT). Измерение уровня экспрессии генов проводили в двух независимых экспериментах, каждый в трёх технических повторах.

Секвенирование образцов МПРА

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.

Анализ данных секвенирования

Данные секвенирования образцов МПРА были проанализированы с помощью собственной программы MPRAdecoder [505], написанной на языке Python. MPRAdecoder находит в прочтениях последовательности индекса, штрихкода и, для образцов картирования, мутации. Затем программа рассчитывает количество прочтений в каждом повторе для каждого штрихкода и конвертирует их в количество прочтений на миллион прочтений повтора. В анализ берутся только штрихкоды длиной от 16 до 20 н., которые имеют не менее трёх прочтений в каждом повторе нормирования. Штрихкоды, которые отличаются друг от друга заменой, инсерцией или делецией одного или двух нуклеотидов, считаются мутантными вариантами одного и того же уникального штрихкода. Количество прочтений уникального штрихкода и его мутантных вариантов суммируется в каждом повторе. Штрихкоды, которые могут быть мутантными вариантами нескольких уникальных штрихкодов, не берутся в дальнейший анализ. 90% прочтений уникального штрихкода в образце картирования должны быть ассоциированы с одной уникальной мутацией, в противном случае они считаются прочтениями химерных продуктов ПЦР и не берутся в дальнейший анализ.

Далее для каждого уникального штрихкода рассчитывается среднее количество прочтений для образцов нормирования, экспрессии и картирования. Уровень экспрессии рассчитывается делением среднего количества прочтений в образце экспрессии на среднее количество прочтений в образце нормирования. Затем рассчитывается логарифмическое значение ($\log_2$) уровня экспрессии. Из полученных значений вычитается логарифмическое значение уровня экспрессии мутации WT, а затем делится на значение логарифмическое значение для мутации ΔC . Благодаря этому полученные значения нормируются таким образом, что во всех библиотеках уровень экспрессии конструкций с мутацией WT равен 0, а ΔC 1.

Программа MPRAdecoder и документация к ней доступны по ссылке <https://github.com/Code-master2020/MPRAdecoder>.

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 FACSCanto II (Becton Dickinson) или NovoCyte (Agilent).

Анализ люминесценции

Для того, чтобы измерить активность люциферазы NanoLuc, трансфицированные клетки HEK293T и CHO собирали через 48 часов после трансфекции. Для этого 800 мкл лизирующего буфера (1×PBS (140 mM NaCl, 2.7 mM KCl, 10 mM Na₂HPO₄, 1.8 mM K₂HPO₄, pH = 7.4), 0.5% Triton X-100) добавляли в каждую лунку 12-тилуночного планшета (3.5 см²) и инкубировали при комнатной температуре 15 минут. Затем 5 мкл клеточного лизата смешивали с 95 мкл лизирующего буфера и 50 мкл раствора субстрата coelenterazine h (NanoLight Technology) в лунках 96-тилуночного белого планшета (SPL Life Sciences). Биолуминисцентный сигнал образца измеряли люминометром Luminoskan Microplate (Thermo Fisher Scientific).

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. У непротрансписированных транскриптов использованных в данной работе плазмидных библиотек для МПРА этот фрагмент пре-мРНК включал в себя СПА, сайт полиаденилирования и район до 56 н. после СПА. У непротрансписированных транскриптов терминатора SV40 early этот район включал в себя СПА, сайт полиаденилирования и район до 41 н. после СПА. У непротрансписированных транскриптов терминатора гена

bGH этот район включал в себя СПА, сайт полиаденилирования и район до 44 н. после СПА.

Для расчёта МСЭ района пре-мРНК, включающего в себя штрихкод, был выбран фрагмент пре-мРНК, расположенный в районе -92..-32 н. выше СПА. Он включает в себя 33 н. до штрихкода, непосредственно штрихкод и 10 н. после него.

МСЭ вторичной структуры РНК была предсказана с помощью программы RNAfold пакета ViennaRNA [506]. Визуализация вторичной структуры РНК была проведена с помощью одноименного интернет-сервиса (<http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi>). Цвета нуклеотидов отражают вероятность нахождения в указанном (спаренном или неспаренном) состоянии. Холодные цвета соответствуют низкой вероятности, а тёплые – высокой.

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/>). Тренировочные данные составляли 91% от общего объёма данных. Для оценки качества работы моделей использовалась кросс-валидация по 10 блокам. Все модели используют технику градиентного бустинга над решающими деревьями (CatBoost, LightGBM, XGBoost). Некоторые модели также используют байесовскую линейную регрессию. Для оптимизации гиперпараметров использовались TPESampler из библиотеки Optuna [507], количество эпох равнялось 50-ти. Целью оптимизации было повышение метрики R^2 (коэффициента детерминации). Для повышения итогового качества предсказаний взвешенные предсказания 2-4 моделей суммировались с помощью VotingRegressor.

Оптимальные веса для каждой модели были подобраны с использованием TPESampler из библиотеки Optuna, количество эпох равнялось 10-ти. Целью оптимизации было повышение метрики R^2 . Для всех полученных в данной работе моделей метрики R^2 и MSE (среднеквадратичная ошибка) имели очень близкие значения для тренировочных и тестовой выборок. На последнем этапе модели были обучены на всём объёме данных – финализированы.

Модель, предсказывающая влияние на уровень экспрессии репортёра последовательности штрихкода, была обучена на данных из библиотеки 49-56, а также на данных для штрихкодов, ассоциированных с мутацией WT, из всех других библиотек. Каждая последовательность штрихкода была охарактеризована с помощью 1438 количественных признаков, значения которых могли варьировать от 0 до 1. Они включали в себя наличие одного из четырёх нуклеотидов в каждой из 18 позиций штрихкода, долю каждого нуклеотида в последовательности штрихкода, частоту встречаемости всех возможных k-меров длиной от 2 до 5 н., и число 1.073, возведённое в степень значения МСЭ. С помощью полученной модели было предсказано влияние штрихкодов на уровень экспрессии репортёров. Значения, предсказанные моделью для штрихкодов, были вычтены из данных.

Модели, предсказывающие уровень зрелой мРНК на основе последовательности после СПА, были обучены на данных, рассчитанных MPRAdecoder, из которых были вычтены значения, предсказанные моделью для штрихкодов. Если последовательности мутации соответствовало несколько штрихкодов в данной плазмидной библиотеке, значения уровней экспрессии для них усреднялись. Каждая последовательность мутации была охарактеризована с помощью 1397 количественных признаков, значения которых могли варьировать от 0 до 1. Они включали в себя наличие одного из четырёх нуклеотидов в каждой из 8 позиций мутации, долю каждого нуклеотида в последовательности штрихкода, частоту встречаемости всех возможных k-меров длиной от 2 до 5 н., и число 1.09543, возведённое в степень

значения МСЭ. На данных об уровнях зрелой мРНК гена eGFP в 8 исследованных плазмидных библиотеках (всех, кроме 49-56) были обучены 8 моделей. На основе этих моделей была разработана финальная модель DoSIA (Downstream Sequence element Iterative Analyzer). DoSIA суммирует взвешенные предсказания 8 моделей. Веса соответствуют метрике R^2 для предсказаний каждой модели на тестовой выборке до финализации, поделённой на сумму R^2 на тестовой выборке для всех 8 моделей. DoSIA позволяет предсказать уровень зрелой мРНК при замене района, расположенного с 3 по 38 н. после сайта полиаденилирования, на любую последовательность соответствующей длины.

DoSIA доступна по ссылке <https://github.com/AnnLetiagina/DoSIA>.

Дизайн новых последовательностей с помощью DoSIA

Для предсказания новых последовательностей была использована собственная программа DSEgenerator. Она генерирует регулируемое пользователем количество последовательностей, которые с большой вероятностью будут обеспечивать высокий или низкий уровень экспрессии репортёра. Последовательность с максимальным или минимальным текущим предсказанным DoSIA уровнем экспрессии сохраняется.

Программа DSEgenerator доступна по ссылке <https://github.com/AnnLetiagina/DoSIA/tree/main/Example>.

Программное обеспечение и оборудование

Анализ данных секвенирования и обучение моделей были проведены на компьютере с процессором Intel® Core™ i3-9100F, оперативной памятью 8 ГБ, 64-разрядной операционной системой Windows 10, версиями Python 3.10.12, Rucaret 3.0.1, Viennarna 2.6.4, scikit-learn 1.2.2 и xgboost 2.0.0.

Результаты

Проведение МПРА

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 [508,509].

Данное исследование направлено на систематический анализ влияния нуклеотидного состава последовательности, расположенной после СПА, на уровень зрелой мРНК вышележащего репортёрного гена eGFP в клетках человека НЕК293Т. 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 11). 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 2-10 (Приложение 2). 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

эквивалентный 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 НЕК293Т, трансфицированные библиотеками для МПРА, собирали для оценки количества ШК в транскриптах eGFP через 48 ч после трансфекции.

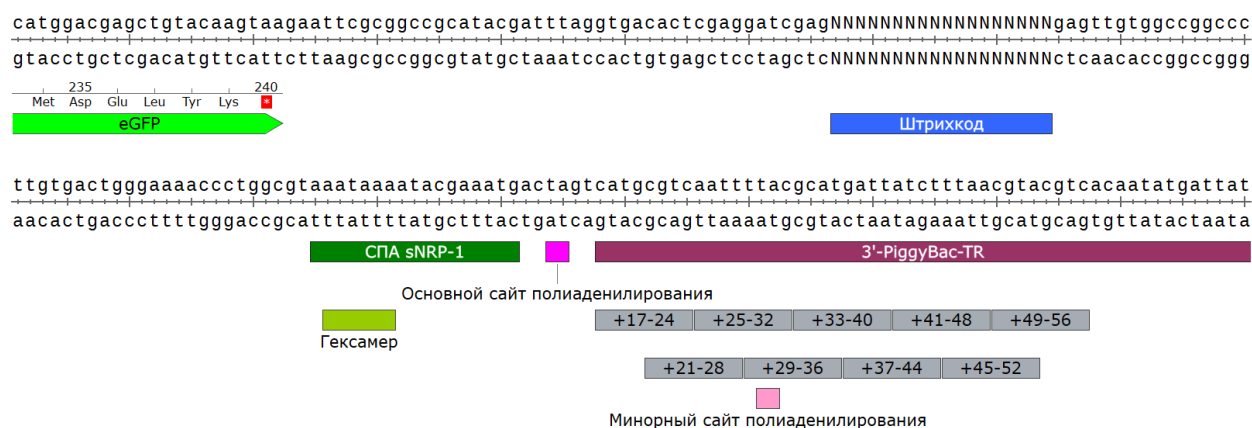


Рисунок 11. Структура библиотек для МПРА (на примере последовательности конструкции дикого типа), использованных для оценки влияния на экспрессию репортёрного гена eGFP последовательностей, расположенных после СПА. Мутации изображены как серые прямоугольники. СПА – сигнал полиаденилирования (гексамер ААТААА); eGFP – 3'-конец кодирующей части гена улучшенного зелёного флуоресцентного белка; sNRP – растворимый нейропилин; TR – терминальный повтор.

Оптимизация условий ПЦР позволяет снизить долю химерных продуктов

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 12). 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 переменных мотива (ШК и мутацию в случае МПРА), разделенных константной областью, часто приводит к образованию нежелательных химерных молекул, в количестве от 5.4 до 30% [510–518].

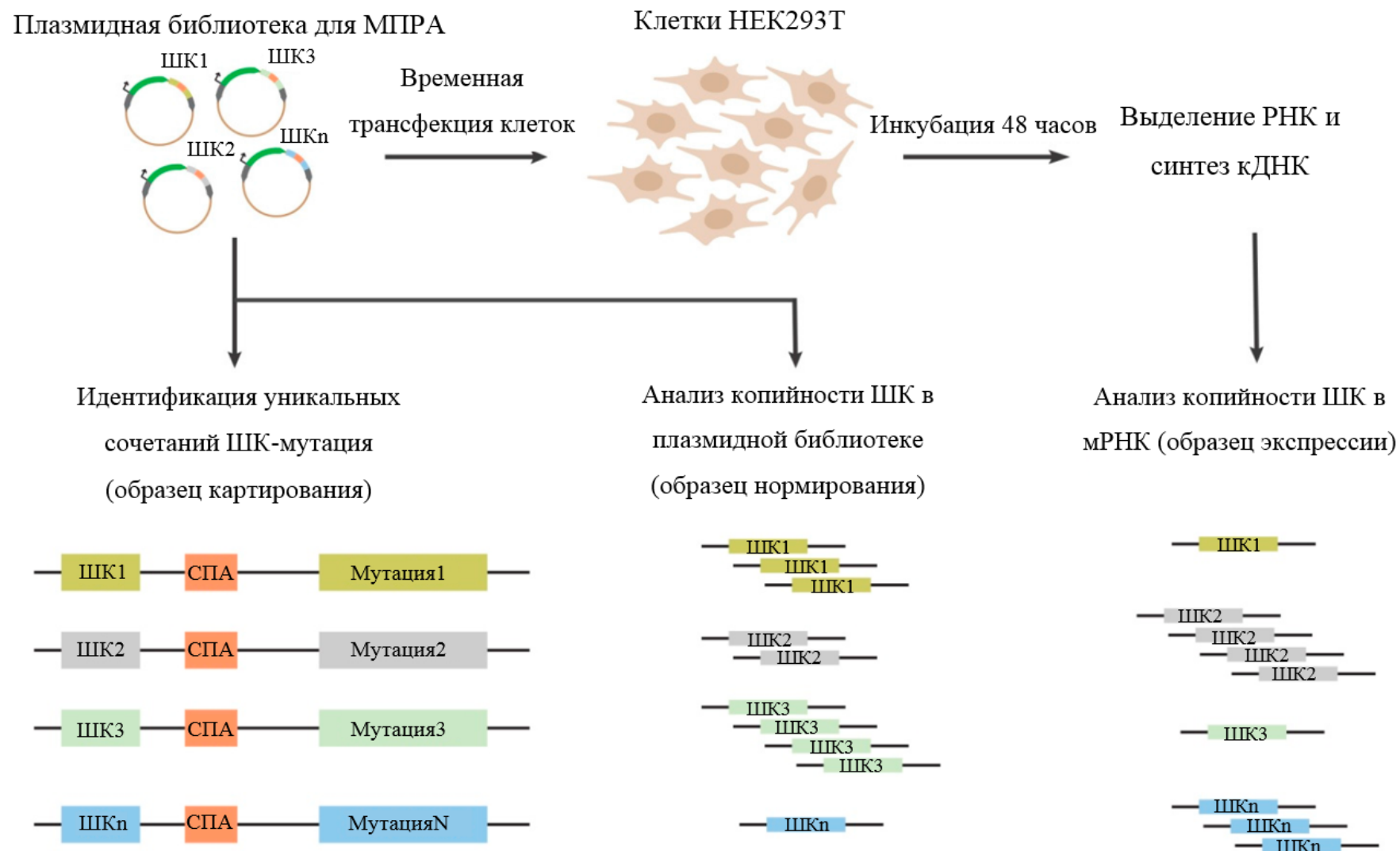


Рисунок 12. Схема экспериментальных этапов приготовления образцов МПРА для последующего высокопроизводительного секвенирования в соответствии с [519]. ШК – штрихкод; СПА – сигнал полиаденилирования.

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 удлинённых праймеров к гетерологичной целевой последовательности во время ПЦР [520]. Неполное удлинение нитей ДНК, предположительно, является следствием остановки ДНК-полимеразы на матрице или её преждевременной терминации. Репликация гетерологичного дуплекса приводит к образованию химерных продуктов ПЦР, состоящих из двух искусственно комбинированных последовательностей [521] (Рисунок 13).

Неполная амплификация в результате паузирования ДНК-полимеразы или её преждевременной терминации

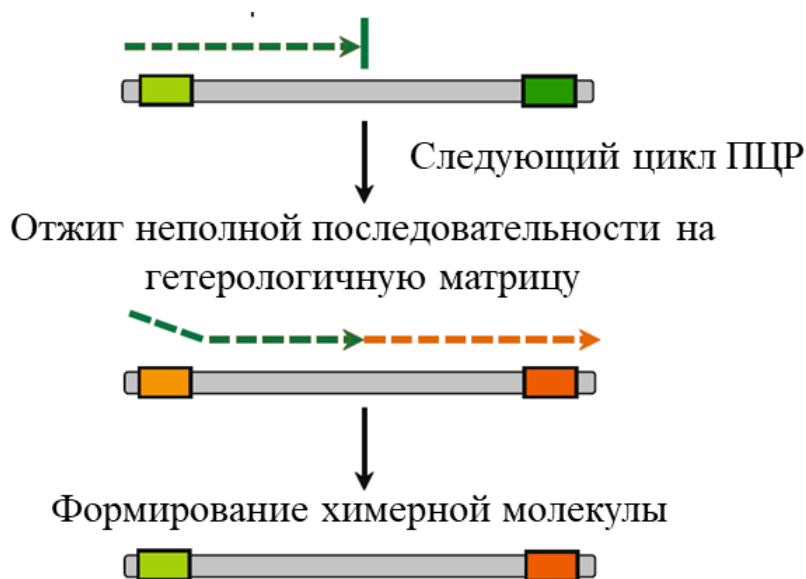


Рисунок 13. Схема формирования химерных молекул ДНК в соответствии с [522].

Частота образования химерных молекул зависит от длины и сходства последовательностей совместно амплифицируемых молекул ДНК [513]. 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

[523,524]. Однако считается, что основной причиной образования химерных молекул является присутствие разных матриц ДНК в одной реакционной смеси ПЦР [510,525]. Таким образом, использование метода эмульсионной ПЦР (эПЦР), обеспечивающего простое физическое разделение молекул ДНК-матрицы за счёт использования эмульсии "вода в масле" [526], представляется прекрасным решением этой проблемы.

Чтобы оценить частоту образования химерных продуктов в ходе ПЦР, была создана простая система, состоящая из двух очень похожих плазмидных конструкций, каждая из которых содержала уникальный ШК длиной 18 п.н. и уникальную восьмибуквенную мутацию в районе +25-32 п.н. после СПА, разделенные константной областью в 71 п.н. (далее плазида №1 и плазида №2). Плазмиды смешивали в эквимольном соотношении и использовали в качестве матрицы для амплификации фрагментов ШК-мутация длиной 236 п.н. с помощью одного раунда традиционной ПЦР. Для выявления наличия химерных комбинаций ШК-мутация в этих продуктах ПЦР были сконструированы пары праймеров BC1/ROI1 и BC2/ROI2 (101-104 в Приложении 2), специфичные для ШК и мутации плазмиды №1 и плазмиды №2, соответственно (Рисунок 14). В тестовой ПЦР для выявления химерных продуктов были использованы пары праймеров BC1/ROI2 и BC2/ROI1. Чтобы исключить образование химерных продуктов в ходе тестовой ПЦР, мы оптимизировали параметры анализа, используя смесь матриц плазмиды №1 и плазмиды №2. В результате, 12 циклов амплификации при температуре отжига 60°C, но не 14 циклов амплификации при 55°C, привели к отсутствию ложноположительных результатов.

Затем фрагменты ШК-мутация были амплифицированы с помощью эПЦР. Мы использовали набор Micellula DNA Emulsion & Purification Kit (EUR_x) и следовали рекомендациям производителя относительно параметров эПЦР (количество мицелл, количество ДНК-матрицы, состав буфера, очистка продуктов эПЦР и т.д.). В частности, мы готовили реакционные смеси

объёмом 5 мкл, содержащие 10^9 - 10^{10} мицелл и 2×10^8 или 2×10^9 молекул пДНК (1 нг или 10 нг смеси плазмиды №1 и плазмиды №2 длиной 4349 п.н. в эквимольном соотношении).

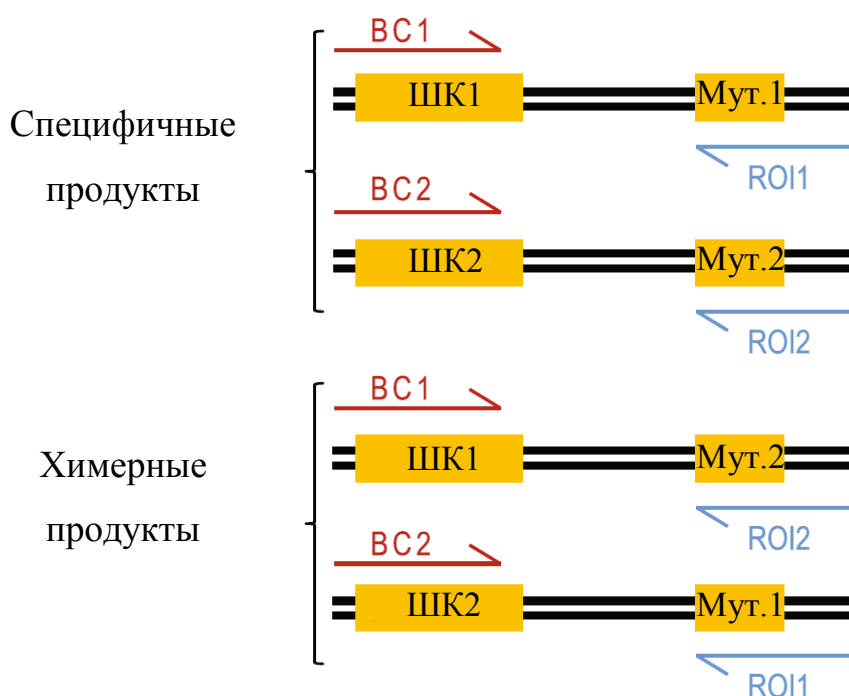


Рисунок 14. Структура всех возможных продуктов ПЦР для системы с двумя плазмидами в качестве матрицы в соответствии с [522]. Красные и синие горизонтальные стрелки обозначают позиции специфических праймеров, использованных в тестовой ПЦР. Мут. – мутация; ШК - штрихкод.

После 25 циклов однораундовой эПЦР эмульсии разбивали 2-бутанолом и очищали продукты реакции. Электрофорез в агарозном геле подтвердил наличие ожидаемых фрагментов ДНК длиной 236 п.н. в образцах "1 нг" и "10 нг" эПЦР (Рисунок 15А). Однако тестовая ПЦР на очищенных продуктах эПЦР, даже полученных с 1 нг матрицы смеси плазмиды №1 и плазмиды №2, выявила химерные молекулы ШК-мутация (Рисунок 15Б).

Чтобы уменьшить образование химерных молекул ШК-мутация в ходе эПЦР, было уменьшено исходное количество матричной ДНК и амплификация была проведена двумя последовательными раундами по 15 и 20 циклов, поскольку одиночные реакции с большим числом циклов могут исчерпать ресурсы индивидуальных мицелл.

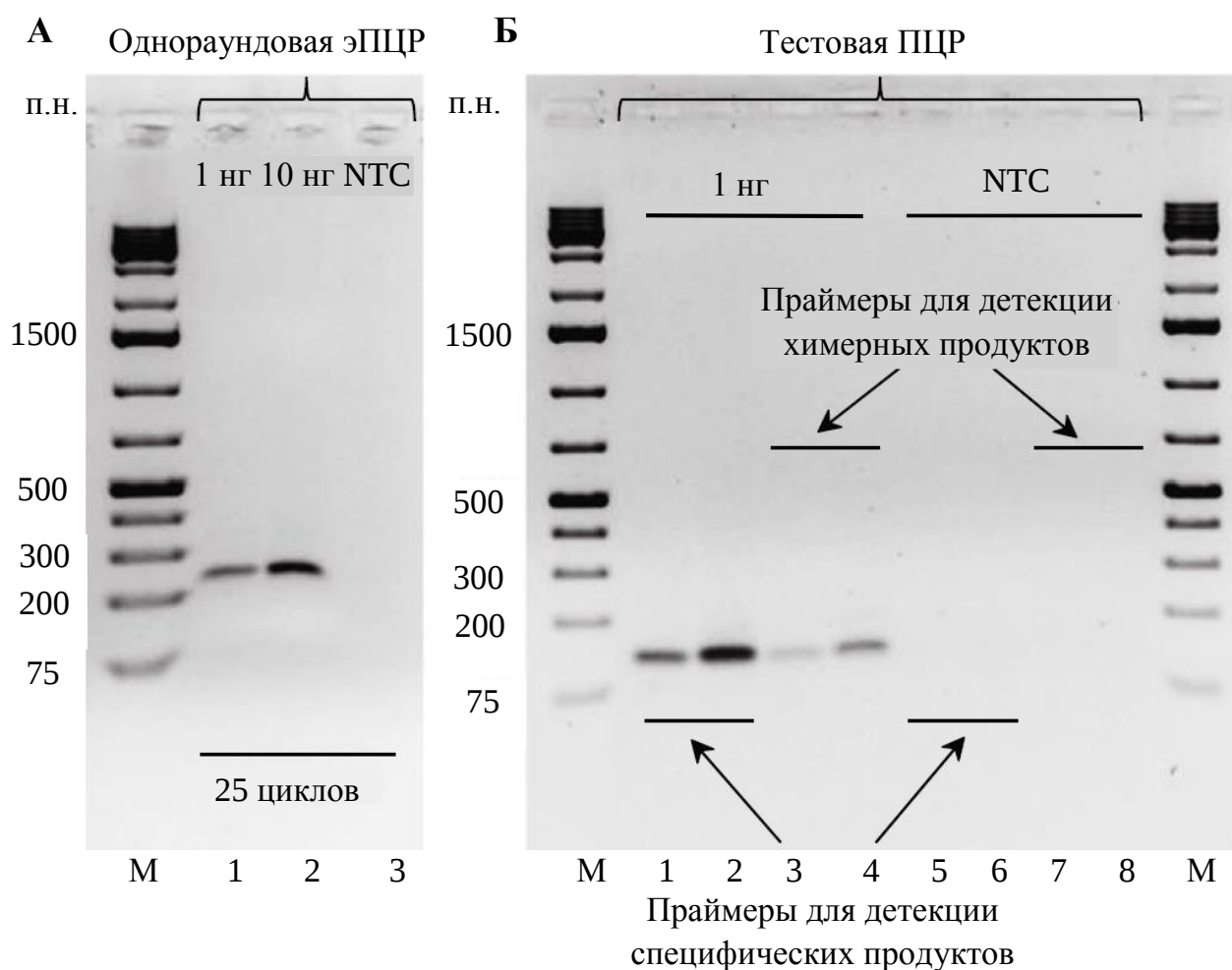


Рисунок 15. Однораундовая эПЦР не подавляет образование химерных молекул с матрицы из смеси двух плазмид в соответствии с [522]. **А)** Анализ продуктов однораундовой эПЦР с помощью электрофореза в агарозном геле, полученных с использованием указанных количеств эквимольных смесей плазмиды №1 и плазмиды №2 (дорожки 1-2) или воды (контроль без матрицы, NTC; дорожка 3) в качестве матрицы и праймеров 20/29 (Приложение 2). **Б)** Анализ продуктов тестовой ПЦР с помощью электрофореза в агарозном геле, полученных с использованием 1/100 части очищенного продукта однораундовой эПЦР "1 нг" (дорожки 1-4) или воды (контроль без матрицы, NTC; дорожки 5-8) в качестве матрицы и следующих праймеров: BC1/ROI1 (дорожки 1 и 5), BC2/ROI2 (дорожки 2 и 6), BC1/ROI2 (дорожки 3 и 7) и BC2/ROI1 (дорожки 4 и 8). М – маркёр GeneRuler 1 kb Plus (Thermo Fisher Scientific).

50 мкл реакций первого раунда эПЦР содержали около 10^9 - 10^{10} мицелл и 2×10^6 , 2×10^7 или 2×10^8 молекул плазмидной ДНК (10 пг, 100 пг или 1 нг эквимольной смеси плазмиды №1 и плазмиды №2). После 15 циклов первого

раунда эПЦР эмульсии разбивали, продукты чистили и анализировали с помощью электрофореза в агарозном геле.

После первого раунда эПЦР в геле не наблюдалось полос (Рисунок 16), что, скорее всего, объясняется малым количеством ДНК-матрицы в сочетании с малым числом циклов амплификации.

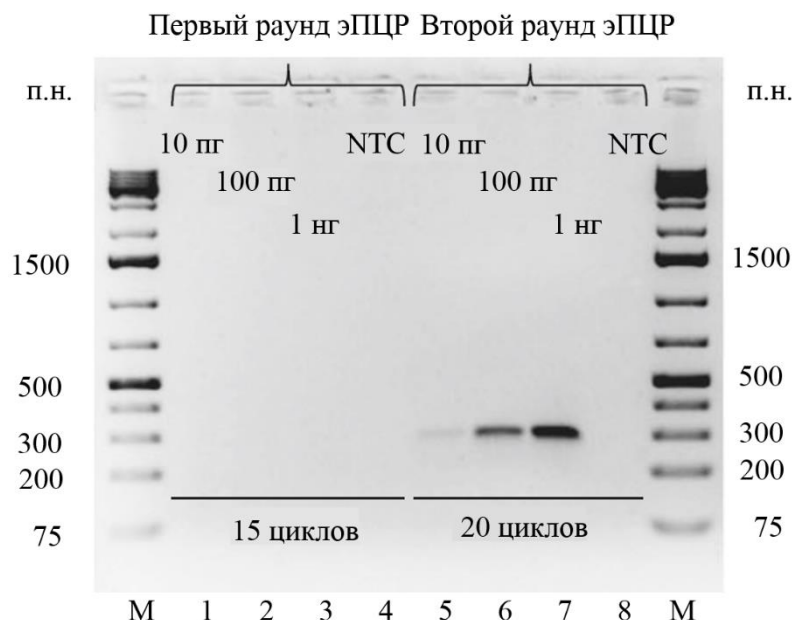


Рисунок 16. Анализ продуктов первого и второго раунда эПЦР с помощью электрофореза в агарозном геле. Взято из [522]. Образцы эПЦР первого раунда были получены с использованием указанных количеств эквимольной смеси плазмиды №1 и плазмиды №2 (дорожки 1-3) или воды (контроль без матрицы, NTC; дорожка 4) в качестве матрицы и праймеров 20 и 29 (Приложение 2). Продукты эПЦР второго раунда были получены с использованием 1/100 части очищенных образцов эПЦР первого раунда (дорожки 5-7) или воды (контроль без матрицы, NTC; дорожка 8) в качестве матрицы и праймеров 20 и 29 (Приложение 2). М – маркер GeneRuler 1 kb Plus (Thermo Fisher Scientific).

Затем был проведён второй раунд эПЦР в объёме 50 мкл, содержащем около 10^9 - 10^{10} мицелл и 1/100 часть (0.5 мкл) очищенных продуктов первого раунда эПЦР. После 20 циклов второго раунда эПЦР продукты были очищены из смесей «10 пг», «100 пг» и «1 нг». Электрофорез в агарозном геле продуктов второго раунда эПЦР выявил четкую прямую зависимость между количеством фрагментов ДНК длиной 289 п.н. и количеством молекул матричной ДНК, использованных для первого раунда эПЦР (Рисунок 16). Примечательно, что

тестовая ПЦР, проведенная с очищенными продуктами «10 пг», но не «100 пг» второго раунда эПЦР, не выявила химерных молекул ШК-мутация (Рисунок 17). Таким образом, мы пришли к выводу, что исходное количество ДНК-матрицы должно быть в $\sim 10^3$ - 10^4 раза меньше, чем предполагаемое общее количество мицелл в эмульсии, чтобы минимизировать вероятность включения двух или более молекул матрицы в одну мицеллу и предотвратить образование химерных продуктов.

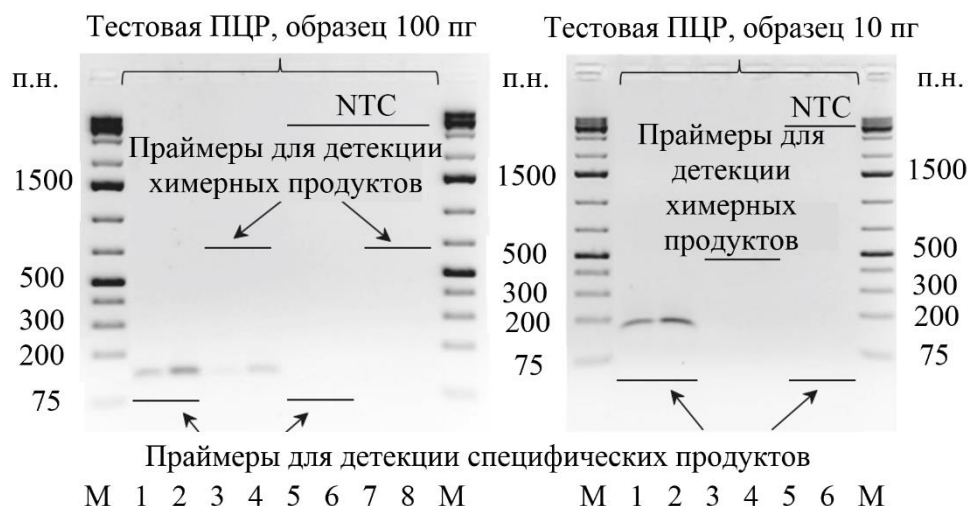


Рисунок 17. Оптимизированная двухраундовая эПЦР эффективно предотвращает образование химерных молекул с матрицы из смеси двух плазмид в соответствии с [522]. **А)** Анализ продуктов тестовой ПЦР с использованием 1/100 объема очищенного образца «100 пг» второго раунда эПЦР (дорожки 1-4) или воды (контроль без матрицы, NTC; дорожки 5-8) в качестве матрицы и следующих праймеров: BC1/ROI1 (дорожки 1 и 5), BC2/ROI2 (дорожки 2 и 6), BC1/ROI2 (дорожки 3 и 7) и BC2/ROI1 (дорожки 4 и 8) с помощью электрофореза в агарозном геле. **Б)** Анализ продуктов тестовой ПЦР с использованием 1/100 объема очищенного образца «10 пг» второго раунда эПЦР (дорожки 1-4) или воды (контроль без матрицы, NTC; дорожки 5-6) в качестве матрицы и следующих праймеров: BC1/ROI1 (дорожки 1 и 5), BC2/ROI2 (дорожки 2 и 6), BC1/ROI2 (дорожка 3) и BC2/ROI1 (дорожка 4). М – маркер GeneRuler 1 kb Plus (Thermo Fisher Scientific).

Для точного измерения доли химерных молекул ШК-мутация, присутствующих в продукте «10 пг» второго раунда эПЦР, он был проанализирован с помощью секвенирования на платформе Illumina MiSeq.

Было выявлено в среднем 1.51% химерных продуктов. Далее была предпринята попытка ещё снизить долю химерных продуктов путем дополнительной оптимизации условий эПЦР. Во-первых, объём ДНК-матрицы, используемой во втором раунде эПЦР, был уменьшен с 0.5 мкл до 0.3 мкл. Во-вторых, количество циклов амплификации во втором раунде эПЦР было уменьшено с 20 до 18. Анализ продуктов амплификации с помощью секвенирования на платформе Illumina показал, что оба условия существенно снижают долю химерных молекул ШК-мутация. В частности, образцы второго раунда эПЦР, полученные с уменьшенным количеством матрицы ДНК и сокращёнными циклами амплификации, содержали в среднем 0.66 и 0.22 % химерных продуктов, соответственно. В целом полученные результаты свидетельствуют о том, что синтез химерных молекул ДНК в ходе амплификации фрагментов ШК-мутация из смеси двух разных матриц может быть подавлен до незначительного уровня с помощью двухраундовой эПЦР.

При совместной амплификации только двух вариантов плазмид частота образования химерных молекул может быть недооценена, поскольку в некоторых мицеллах могут находиться идентичные молекулы-матрицы (это в первую очередь относится к мицеллам, в которые попали две плазмиды). Однако при совместной амплификации большого числа гомологичных последовательностей ДНК вероятность включения одинаковых молекул-матриц в одну мицеллу крайне мала. Поэтому доля химерных молекул в образцах эПЦР, амплифицированных с матрицы из смеси множества плазмид, может быть выше, чем та, которая была обнаружена для описанной выше системы с матрицей из смеси двух плазмид. Чтобы проверить это, мы применили двухраундовую эПЦР для амплификации фрагментов ШК-мутация из двух библиотек для МПРА с высоким разнообразием – 25-32 и 37-44. Плазмиды обеих библиотек содержат 18-буквенный ШК и 8-буквенную мутацию, которые разделены консервативной последовательностью длиной 71 п.н. (библиотека 25-32) или 83 п.н. (библиотека 37-44). Сначала были амплифицированы фрагменты ШК-мутация библиотеки 25-32, используя

оптимальные параметры эПЦР, определенные для системы с матрицей из смеси двух плазмид (10 пг пДНК в первом раунде эПЦР, 0.5 мкл очищенных продуктов первого раунда эПЦР в качестве матрицы и 18 циклов амплификации во втором раунде эПЦР).

На основании данных секвенирования на платформе Illumina MiSeq был определён набор истинных последовательностей ШК, присутствующих в плазмидной библиотеке, как это описано в [7]. Кратко, из прочтений были выделены последовательности ШК и мутации. Затем были идентифицированы мутантные версии ШК (возникшие в результате ошибок ПЦР и/или NGS), содержащие до 2 нуклеотидных замен и обнаруженные в малом количестве прочтений, их ассоциировали с соответствующими подлинными ШК. Для последующего анализа оставляли только подлинные ШК, найденные не менее чем в двух повторах секвенирования плазмидной библиотеки. Затем для каждого подлинного ШК считали связанные с ним последовательности мутаций. Последовательность мутации, найденная более чем в половине прочтений, несущих подлинный ШК (с частотой >0.5), считалась подлинной мутацией. Последовательности мутаций, отличающиеся на 1 нуклеотид от подлинной мутации, рассматривались как вариации мутации, возникшие из-за ошибок ПЦР и/или NGS, в то время как все остальные последовательности мутаций считались результатом химерной ПЦР. Данный анализ был проведён при помощи собственной программы MPRAdecoder, описанной в [505].

В результате проведённого анализа было обнаружено, что химерные молекулы ШК-мутация образуются со средней частотой 0.57 %, что более чем в два раза выше, чем в системе с матрицей из смеси двух плазмид. Для дальнейшей оптимизации условий эПЦР была увеличена продолжительность этапа элонгации с 10 до 30 сек. Фрагменты ШК-мутация библиотек 25-32 и 37-44 были амплифицированы с изменением этого параметра эПЦР и проанализированы с помощью секвенирования на платформе Illumina MiSeq. Анализ данных секвенирования показал высокую воспроизводимость

измерений как между повторами, так и между библиотеками, а также то, что в среднем образование химерных продуктов было снижено почти в 2 раза, до 0.30 %.

Далее, чтобы напрямую сравнить эффективность эмульсионной и обычной ПЦР, мы повторили амплификацию фрагментов ШК-мутация обеих плазмидных библиотек с помощью обычной ПЦР с теми же настройками, которые использовались для эПЦР (10 пг матричной ДНК в первом раунде ПЦР, 0.5 мкл очищенных продуктов первого раунда ПЦР в качестве матрицы и 18 циклов амплификации во втором раунде ПЦР, время элонгации 30 с в обоих раундах ПЦР). Неожиданно, анализ данных высокопроизводительного секвенирования показал, что доля химерных молекул ШК-мутация в продуктах обычной ПЦР (в среднем 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. Результаты данной работы были опубликованы [522].

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 12).

Образцы экспрессии и нормирования были подготовлены путем ПЦР-амплификации последовательностей ШК с использованием кДНК, полученной из трансфицированных клеток НЕК293Т, и ДНК плазмидной библиотеки в качестве матрицы, соответственно (Рисунок 12). 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

Анализ данных высокопроизводительного секвенирования был проведён с помощью собственной программы MPRAdecoder [505]. Он показал, что MPRA-библиотеки имеют разное количество уникальных мутаций (Таблица 2). В среднем по библиотекам, 67%, 19% и 14% мутаций были связаны с одной, двумя, тремя и более ШК, соответственно. Следует отметить, что для всех библиотек МПРА была получена высокая корреляция нормированных значений экспрессии между повторами (Рисунок 18). Для мутаций, ассоциированных с двумя и более ШК, уровень экспрессии вычислялся как среднее значение уровней экспрессии ассоциированных с ней ШК.

Таблица 2. Число уникальных мутаций и ассоциированных с ними штрихкодов для каждой библиотеки для МПРА.

Библио-тека	Число мута-ций	Мутации, ассоциированные с 1 ШК		Мутации, ассоциированные с 2 ШК		Мутации, ассоциированные с ≥ 3 ШК	
		Количество	%	Количество	%	Количество	%
17-24	8517	7688	90.3	745	8.7	84	1
21-28	8386	7444	88.8	813	9.7	129	3.1
25-32	46023	20693	45	13252	28.8	12078	26.2
29-36	10019	8693	86.8	999	10	327	3.3
33-40	27025	19060	70.5	5869	21.7	2096	7.8
37-44	33191	18772	56.6	8588	25.9	5831	17.6
41-48	53837	17353	32.2	14623	27.2	21861	40.6
45-52	25498	17917	70.3	5537	21.7	2044	8
49-56	17870	11862	66.4	3478	19.5	2530	14.2

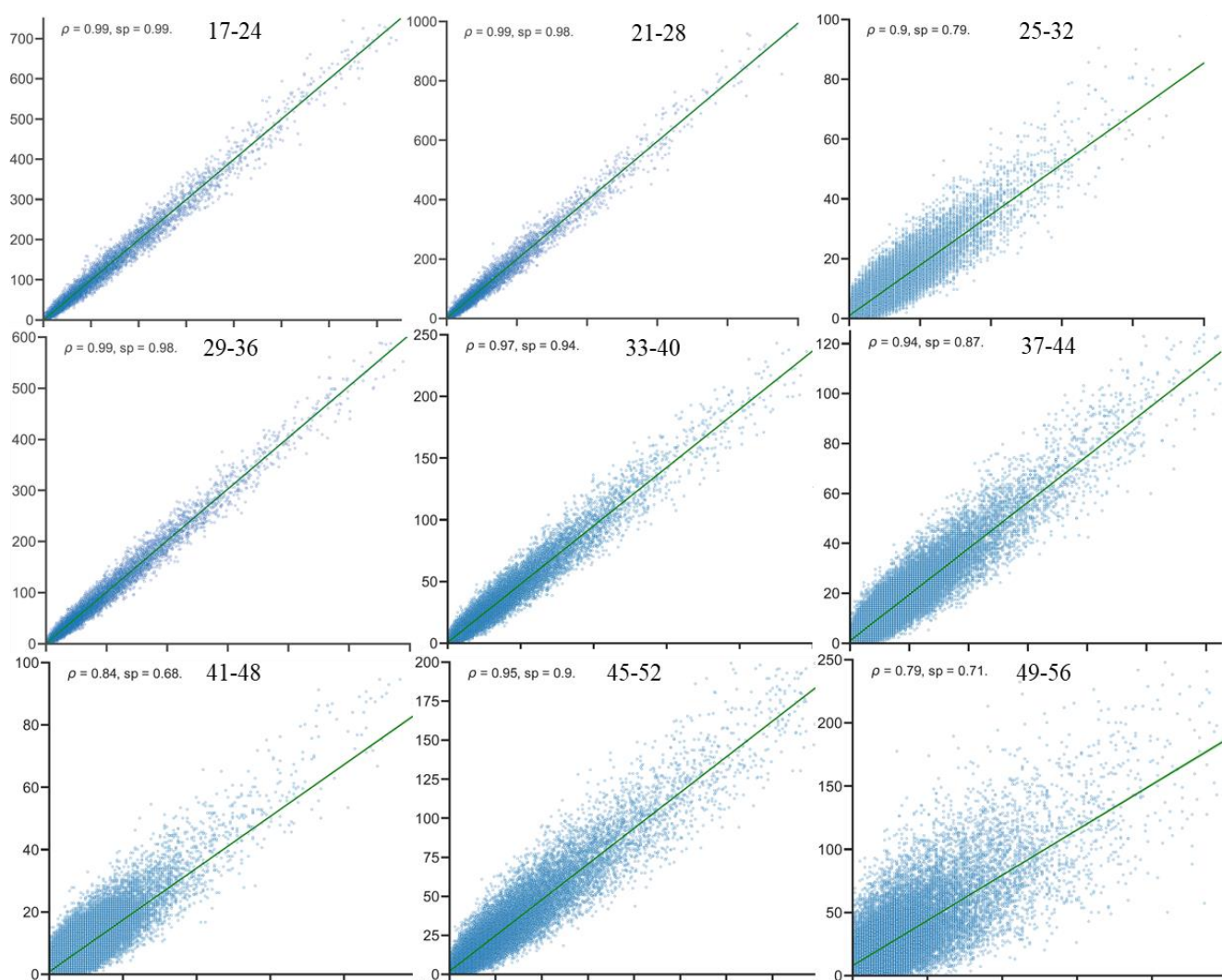


Рисунок 18. Корреляция нормированных уровней экспрессии между повторами библиотек для МПРА. По оси абсцисс указано количество прочтений штрихкода в первом повторе экспрессии, нормированное на миллион прочтений первого повтора экспрессии, по оси ординат – количество прочтений штрихкода во втором повторе экспрессии, нормированное на миллион прочтений второго повтора экспрессии. ρ – коэффициент корреляции Пирсона; sp – коэффициент корреляции Спирмена.

Далее мы оценили количество мутаций в каждой библиотеке для МПРА, которые увеличивают или уменьшают уровень зрелой мРНК eGFP более чем в два раза по сравнению с контрольной конструкцией WT (Таблица 3, Рисунок 19).

Таблица 3. 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. Столбец «>0/<0» отражает, на сколько число последовательностей, повышающих уровень зрелой мРНК eGFP, превышает число последовательностей, понижающих уровень зрелой мРНК eGFP. В нижней строке приведены данные о влиянии последовательностей ШК. МПРА – массовый параллельный репортёрный анализ. Макс. – максимальная; Мин. – минимальная; Ср. – средняя; ШК – штрихкоды; Экспр. – экспрессия (уровень зрелой мРНК eGFP); ρ – коэффициент корреляции Пирсона; σ – стандартное отклонение.

Библиотека	Число последовательностей с уровнем экспр. > 0	Макс. экспр., $\log_2$	Мин. экспр., $\log_2$	Ср. экспр., $\log_2$	σ	=>1, %, $\log_2$	<=-1, %, $\log_2$	> 0 / < 0, $\log_2$	ρ для мутаций с двумя и более ШК
17-24	8513	2.55	-1.6	0.9	0.46	43	0.1	30.4	0.43
21-28	8380	1.96	-1.89	0.57	0.42	15	0.1	11.8	0.63
25-32	45803	3.76	-3.13	0.94	0.87	48	2	6.5	0.47
29-36	10021	3.53	-1.93	1.35	0.64	70	0.05	49.6	0.71
33-40	27017	4.26	-3.02	1.59	0.8	78	0.3	37.3	0.52
37-44	33141	2.38	-2.4	0.45	0.48	11	1	5.1	0.37
41-48	53605	2.17	-2.02	0.07	0.38	0.4	1	1.7	0.19
45-52	25473	1.36	-2.31	0.09	0.29	0.1	0.1	1.9	0.27
49-56	17635	1.81	-2.54	0.004	0.41	0.2	0.6	1.2	0.06
49-56 + WT (ШК)	36735	2.39	-2.86	0.06	0.52	0.02	0.04	1.4	-

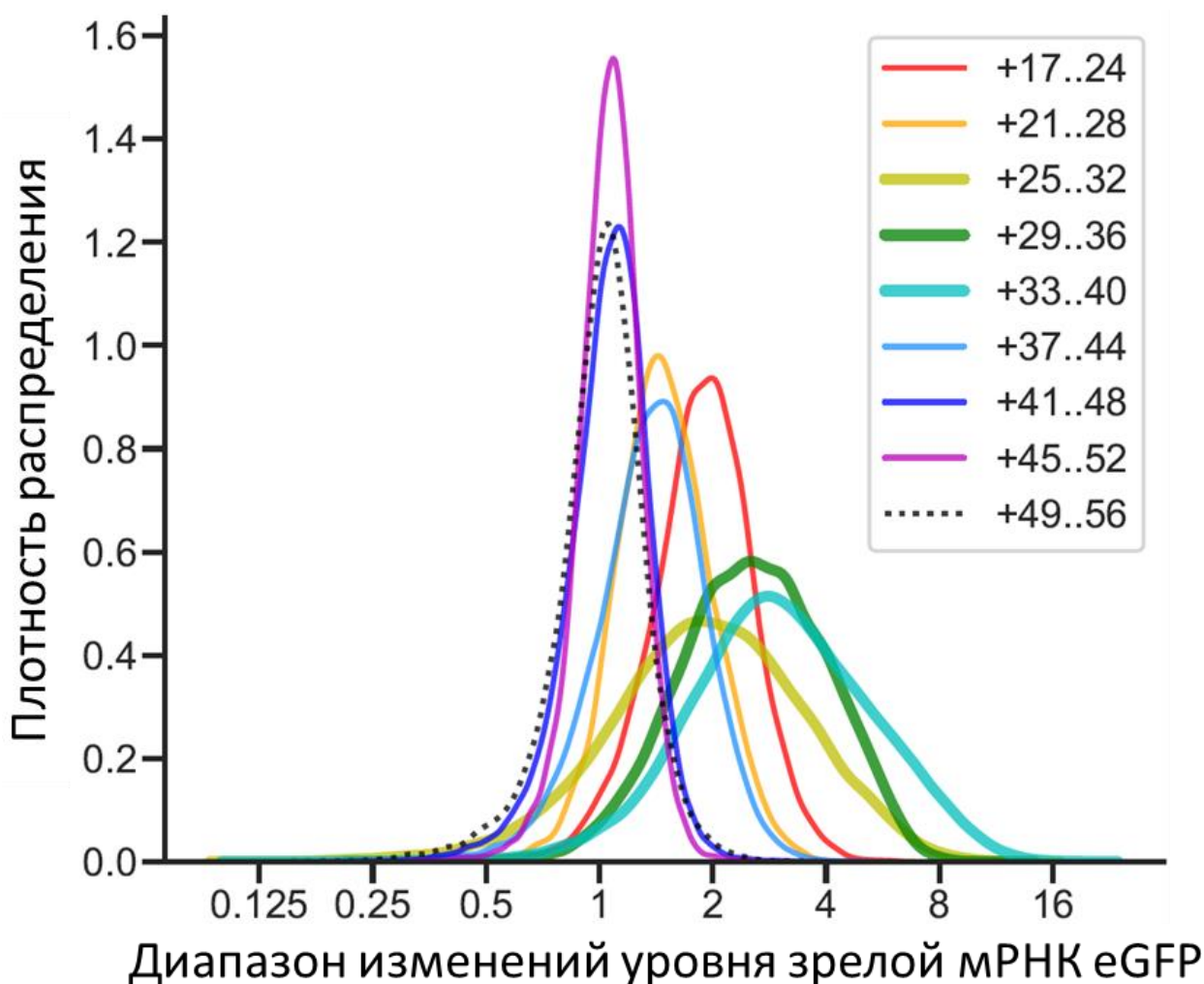


Рисунок 19. Влияние изменения последовательности различных районов, расположенных после СПА, на количество зрелой мРНК eGFP. В верхнем правом углу указано положение варибельной области. СПА – сигнал полиаденилирования.

Наибольшее влияние на уровень зрелой мРНК оказывают последовательности, расположенные в районе +17..+40 н. после СПА

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. 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.

Известно, что последовательности, расположенные выше СПА, могут оказывать влияние на сборку канонического комплекса процессинга 3'-конца пре-мРНК [13,33,62,165]. 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 (Таблица 4). 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 (Таблица 4, Рисунок 20). Это подтверждается низким коэффициентом корреляции Пирсона – 0.06 – для мутаций с двумя и более штрихкодами в библиотеке 49-56 (Таблица 3). Тест Левена также не выявил отличий в величинах дисперсий для этих выборок в библиотеке 49-56 (p-значение = 0.74).

Таблица 4. ШК, ассоциированные с исходной последовательностью после СПА (WT). В таблице указано логарифмическое значение ($\log_2$) нормированного уровня зрелой мРНК eGFP. Макс. – максимальный; Мин. – минимальный; СПА – сигнал полиаденилирования; Ср. – средняя; Экспр. – экспрессия (уровень зрелой мРНК eGFP); σ – стандартное отклонение.

Библио-тека	Число ШК с мутацией WT	Макс. экспр.	Мин. экспр.	Ср. экспр.	σ	U-критерий Манна-Уитни
17-24	11 (0.1%)	0.39	-0.24	0.01	0.2	$2e^{-7}$
21-28	31 (0.3%)	0.88	-0.92	-0.03	0.29	$5e^{-14}$
25-32	98 (0.1%)	1.28	-2.12	-0.04	0.71	$3e^{-28}$
29-36	30 (0.2%)	0.49	-0.7	-0.06	0.31	$1e^{-18}$
33-40	275 (0.7%)	1.8	-2.3	-0.05	0.58	$4e^{-143}$
37-44	5781 (9.2%)	1.83	-2.1	-0.05	0.47	0
41-48	450 (0.3%)	1.23	-1.39	-0.04	0.39	$2e^{-11}$
45-52	45 (0.1%)	0.77	-1.46	-0.04	0.42	0.05
49-56	11 (0.1%)	0.81	-1.41	0	0.52	0.6

Затем был проанализирован нуклеотидный состав 5% мутаций, ассоциированных с наиболее высоким уровнем зрелой мРНК eGFP в каждой библиотеке с использованием интернет-сервиса pLogo [527]. Оказалось, что для мутаций, обеспечивающих высокий уровень экспрессии eGFP, характерно выраженное обогащение тиминот (Рисунок 21). 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 [127,129,130], тогда как мутации с уровнем экспрессии eGFP ниже, чем у исходной последовательности, были обогащены мотивами AGG, GGG, GGC, AGGG, CAGG, GGCC, AGGC.

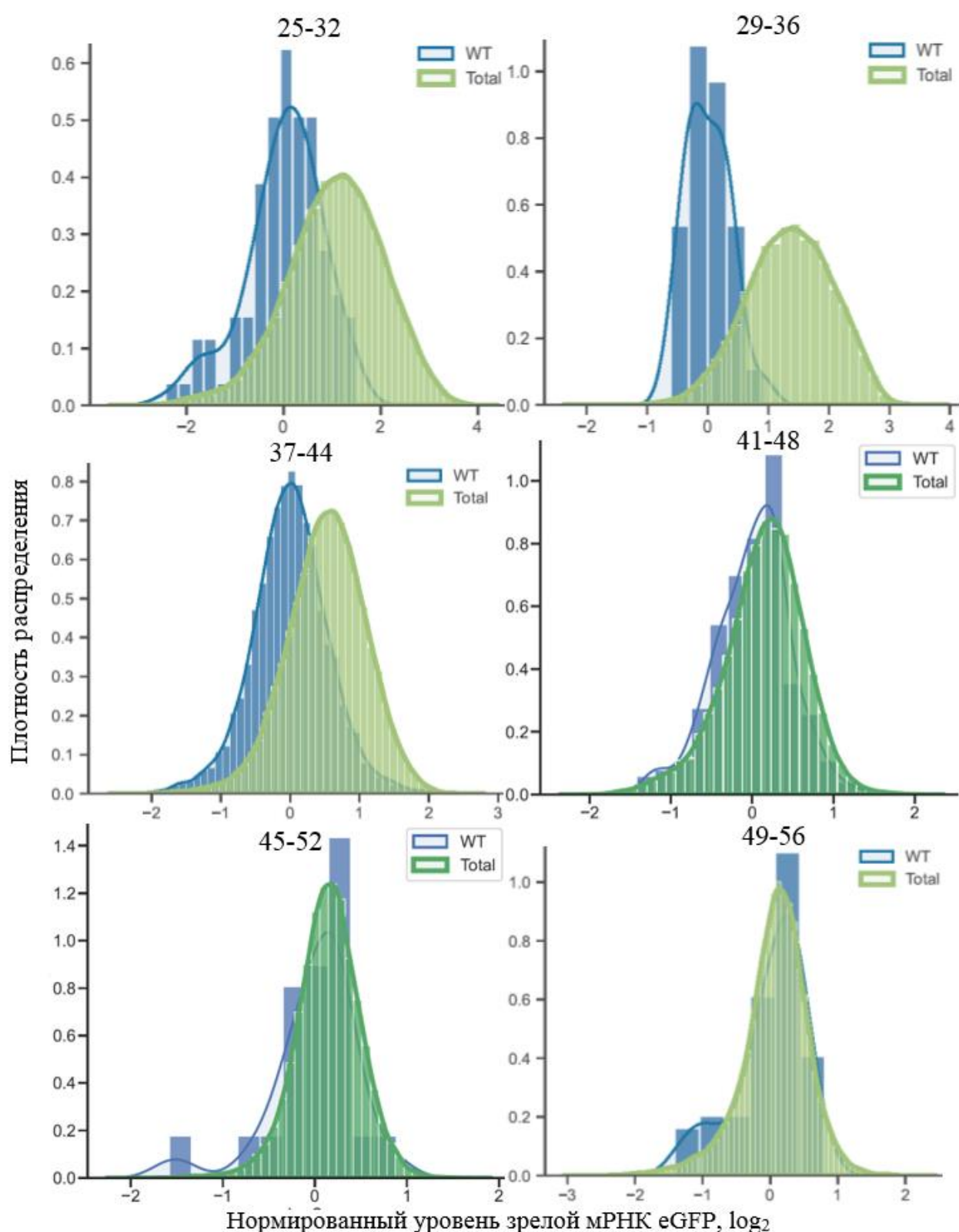


Рисунок 20. Плотности распределения уровней зрелой мРНК eGFP для ШК, ассоциированных с исходной последовательностью после СПА (WT, синий), и мутациями (Total, зелёный) для плазмидных библиотек 25-32, 29-36, 37-44 и 49-56. Изменение последовательности в исследуемых районах после СПА, кроме +49..56, оказывает влияние на уровень зрелой мРНК eGFP.

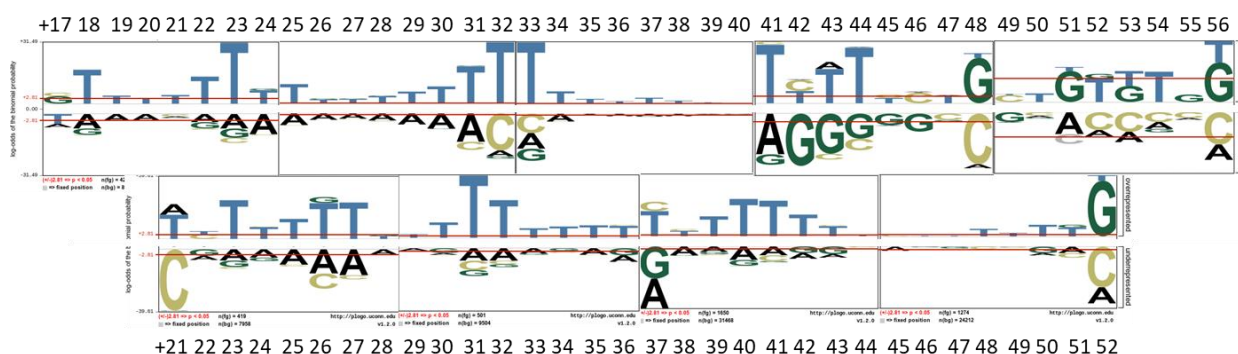


Рисунок 21. Мутации, ассоциированные с высоким уровнем зрелой мРНК eGFP, отличаются выраженным обогащением тиминном. Визуализация позиционной весовой матрицы, полученной с помощью интернет-сервиса pLogo [527]. Пороговый уровень достоверности 0.05 обозначен красными горизонтальными линиями.

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 непротессированных транскриптов гена eGFP этот фрагмент пре-мРНК включал в себя СПА, сайт полиаденилирования и район до 56 н. после СПА. Оказалось, что непротессированные транскрипты исходной последовательности имеют вторичную структуру после СПА (Рисунок 22). При этом делеция цитозина в позиции + 32 н. после СПА, которая, как было показано ранее [8], приводит к повышению уровня зрелой мРНК и белка eGFP в 2 раза, снижает стабильность вторичной структуры в выбранном районе пре-мРНК.

Известно, что СПА, сайт полиаденилирования и расположенный после него DSE являются сайтами связывания множества РНК-связывающих белков, вовлечённых в процессинг 3'-конца пре-мРНК [11,31,60]. При этом большинство РНК-связывающих белков предпочтительно связывают оцРНК [522].

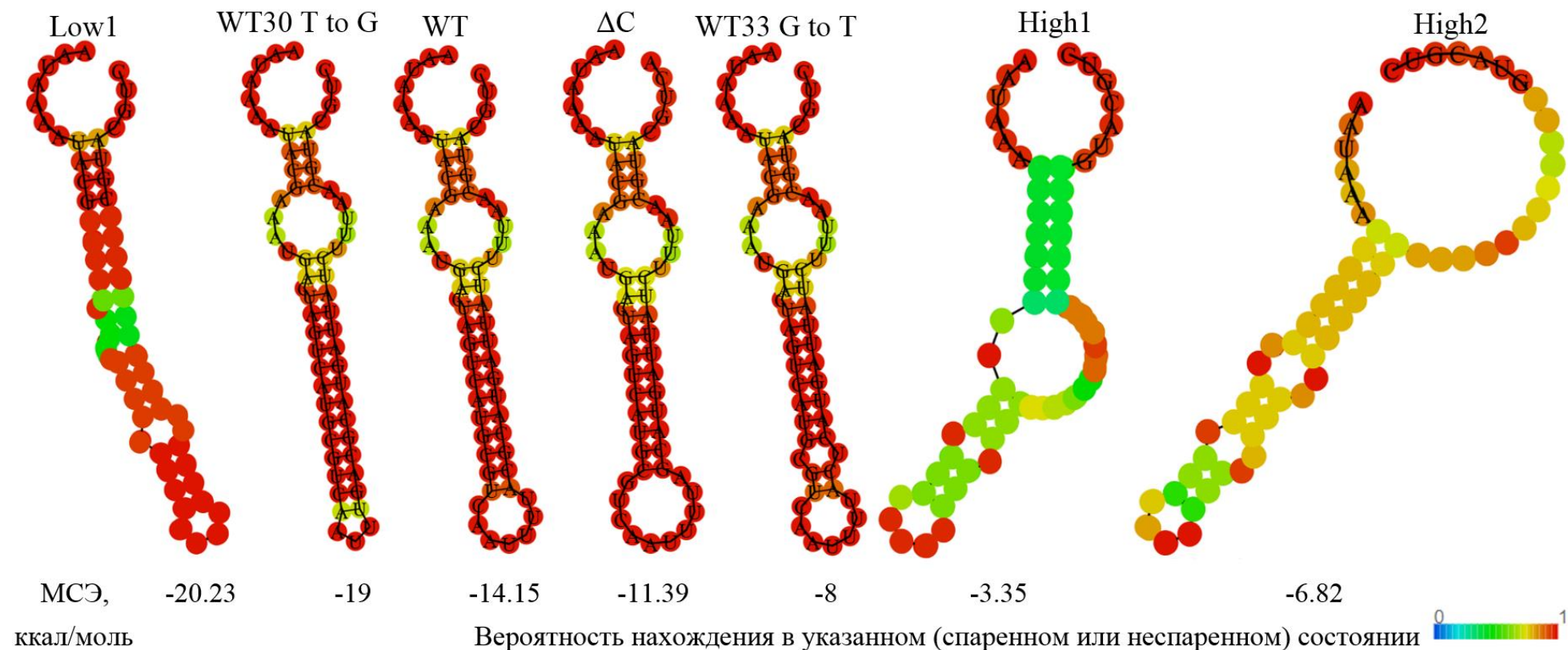


Рисунок 22. Пре-мРНК eGFP образует шпильку в районе после СПА растворимого нейропилина 1. Для ряда последовательностей, исследуемых в данной работе, были визуализированы предсказанные с помощью веб-сервиса RNAfold (<http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi>) вторичные структуры фрагмента пре-мРНК, включающего в себя СПА, сайт полиаденилирования и район до 56 н. после СПА. Цвета нуклеотидов отражают вероятность нахождения в указанном (спаренном или неспаренном) состоянии. Холодные цвета соответствуют низкой вероятности, а тёплые – высокой. MCЭ – минимальная свободная энергия; WT – исходная последовательность; ΔC – делеция C в позиции +32 п. 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 5).

Таблица 5. 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; σ – стандартное отклонение. * – p-значение <0.05, ** – p-значение <0.01, *** – p-значение <0.001, **** – p-значение <0.00001.

Библио-тека	ρ (МСЭ, экспр.)	Макс. МСЭ	Мин. МСЭ	Ср. МСЭ	σ	Уровень достоверности точного критерия Фишера
17-24	0.1****	-1	-16.7	-3	2	0.09
21-28	0.42****	-6.7	-17	-9.3	1.8	$1e^{-31}$
25-32	0.42****	-10.8	-20.7	-12.5	1.6	0
29-36	0.19****	-1.6	-18.4	-5.5	2.5	$2e^{-18}$
33-40	0.17****	-2.4	-15.9	-3.4	1.4	$7e^{-4}$
37-44	0.35****	-4.8	-24.1	-8.1	2.4	$5e^{-213}$
41-48	0.28****	-11.4	-31.4	-15.1	2.3	$3e^{-146}$
45-52	-0.02***	-11.4	-25.4	-13.5	1.3	-
49-56	-0.02**	-11.4	-18.4	-12.7	0.4	-
49-56 + WT (ШК)	0.01*	-5.7	-29.6	-12.2	2.7	-

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 5).

Ранее в ряде работ уже было показано, что повышение стабильности вторичной структуры вблизи СПА в пре-мРНК приводит к снижению уровня экспрессии репортёрного гена [185,186]. 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

По синим графикам на рисунке 19 можно сделать вывод, что последовательности ШК оказывают влияние на уровень зрелой мРНК eGFP наряду с последовательностями мутаций. В библиотеке 49-56, как было показано ранее (Таблица 4, Рисунок 20), последовательности мутаций не оказывают влияния на уровень экспрессии репортёра и наблюдаемая вариабельность уровней экспрессии репортёра может быть обусловлена именно влиянием последовательности ШК. Среднее логарифмическое значение экспрессии ШК, ассоциированных с мутациями WT, сводится к 0 при нормировке данных МПРА, проводимой MPRAdecoder. Для того, чтобы проанализировать влияние штрихкодов, мы объединили данные из библиотеки 49-56 с данными о ШК, ассоциированных с мутацией WT из всех

остальных библиотек (Таблица 3, библиотека «49-56 + WT»). Таким образом, библиотека ШК «49-56 + WT» представляет собой набор данных, позволяющих судить о влиянии последовательности ШК на уровень экспрессии репортёра.

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 3, нижняя строка). 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 5, нижняя строка). 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 отметить, 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» (Таблица 6). С помощью полученной модели было предсказано влияние всех исследованных в данной работе ШК на уровень экспрессии репортёров. Значения, предсказанные моделью для ШК, были вычтены из данных МПРА, обработанных MPRAdecoder.

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 (Таблица 6). Если последовательности мутации соответствовало несколько ШК в данной плазмидной библиотеке, значения уровней экспрессии для них усреднялись. Для всех полученных в данной работе моделей метрики R^2 и MSE имели очень близкие значения для тренировочной и тестовой выборки.

Интересно отметить, что для моделей, предсказывающих уровень зрелой мРНК на основе последовательности района после СПА, метрика R^2 (Таблица 6) имеет схожие значения с коэффициентом корреляции Пирсона для мутаций с двумя и более штрихкодами (Таблица 3, крайний правый столбец) в одноименных библиотеках ($\rho = 0.98$). Мы предполагаем, что оба этих показателя отражают относительную силу влияния каждого исследованного района после СПА на уровень экспрессии репортёрного гена.

Таблица 6. Характеристики моделей, предсказывающих уровень зрелой мРНК в зависимости от последовательности мутации или ШК. ШК – штрихкод; MSE – среднеквадратичная ошибка; R^2 – коэффициент детерминации; test – для тестовой выборки; train – для тренировочной выборки.

Модель для библиотеки	R^2 train	R^2 test	MSE train	MSE test
+17-24	0.42	0.46	0.121	0.117
+21-28	0.59	0.6	0.073	0.077
+25-32	0.52	0.51	0.36	0.386
+29-36	0.67	0.66	0.137	0.152
+33-40	0.54	0.55	0.301	0.273
+37-44	0.4	0.38	0.141	0.14
+41-48	0.27	0.27	0.103	0.098
+45-52	0.28	0.31	0.061	0.055
+49-56 + WT (ШК)	0.09	0.07	0.241	0.264

При этом модель, предсказывающая влияние последовательности ШК на уровень экспрессии репортёра, обладает самым низким R^2 среди всех моделей (0.07 для тестовой выборки), хотя все модели были обучены схожим образом. Мы предполагаем, что это может быть связано с тем, что большая часть последовательностей ШК оказывает очень слабое влияние на уровень экспрессии репортёра.

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.

Сравнение полученной нами модели DoSIA с уже известной APARENT2

Мы провели сравнение производительности DoSIA с APARENT2 [421] – одной из лучших на сегодняшний день моделей для предсказания влияния СПА и окружающих его последовательностей на уровень экспрессии и выбор сайта полиаденилирования гена. Корреляция Пирсона DoSIA с данными проведённого нами МПРА составила 0.77, корреляция Пирсона APARENT2 с данными проведённого нами МПРА составила 0.37. Впрочем, нет ничего удивительного в том, что производительность DoSIA оказалась выше APARENT2, так как она была обучена на подобных данных.

Мы также провели сравнение DoSIA и APARENT2 на опубликованных данных о влиянии разнообразных мутаций в широко используемых терминаторах транскрипции на уровни разрезания (анализ разрезания *in vitro*), зрелой мРНК (анализ защиты от РНКазы или нуклеазы S1) и белка (анализ люминесценции) репортёрного гена (Таблица 7). В анализ были взяты только работы, в которых авторы исследовали мутации, затрагивающие область DSE, проанализировали более 10 последовательностей и провели количественный анализ полученных результатов.

В результате мы обнаружили, что производительность DoSIA немного уступает производительности APARENT2 на данных McDevitt *et al.*, 1986 [500]. Это может быть связано с тем, что APARENT2, в отличие от DoSIA, может учитывать взаимное влияние DSE и окружающих его цис-элементов, принимающих участие в процессинге 3'-конца пре-мРНК. Производительность DoSIA немного превосходит производительность APARENT2 на данных Chou *et al.*, 1994 [501]. Это может быть связано с тем, что DoSIA, в отличие от APARENT2, может напрямую учитывать влияние вторичной структуры терминатора SV40 late [186,504]. Производительность DoSIA значительно превосходит производительность APARENT2 на данных Wu and Alwine, 2004 [186]. Это может быть связано с тем, что авторы этой работы целенаправленно исследовали влияние вторичной структур

Таблица 7. DoSIA и APARENT2 демонстрируют схожую производительность на данных о влиянии разнообразных мутаций на уровни разрезания, зрелой мРНК и белка репортёрного гена. МСЭ – минимальная свободная энергия; РНКазы – рибонуклеаза; APARENT – APA REgression NeT; bGH – бычий гормон роста; DoSIA – Downstream Sequence element Iterative Analyzer; ρ – коэффициент корреляции Пирсона. [†] – р-значение < 0.1, * – р-значение < 0.05, ** – р-значение < 0.01, *** – р-значение < 0.001, ***** – р-значение < 0.00001.

Статья	Исследуемый терминатор	Модельный объект	Метод измерения	Объём выборки	ρ , DoSIA	ρ , DoSIA, МСЭ = 0	ρ , APARENT2
McDevitt <i>et al.</i> , 1986 [500]	SV40 early	Клетки линии HEK293	Анализ защиты от РНКазы	17	0.82 ^{*****}	0.83 ^{*****}	0.84 ^{*****}
				20	0.89 ^{*****}	0.88 ^{*****}	0.95 ^{*****}
Chou <i>et al.</i> , 1994 [501]	SV40 late	Ядерный экстракт клеток HeLa	Анализ разрезания <i>in vitro</i>	16	0.77 ^{***}	0.76 ^{***}	0.74 ^{***}
Wu and Alwine, 2004 [186]	SV40 late	Ядерный экстракт клеток HeLa S3	Анализ разрезания <i>in vitro</i>	13	0.75 ^{**}	0.63 [*]	0.7 ^{**}
		Клетки CV-1	Анализ люминисценции	8	0.86 ^{**}	-	0.76 [*]
Goodwin and Rottman, 1992 [503]	bGH	Клетки COS-1	Анализ защиты от нуклеазы S1	16	0.46 [†]	0.74 ^{**}	0.89 ^{*****}

терминатора SV40 late, и способность DoSIA напрямую учитывать их стабильность дала ей заметное преимущество. Производительность DoSIA значительно уступает производительности APARENT2 на данных Goodwin and Rottman, 1992 [503]. Это может быть связано с уникальной особенностью терминатора гена *bGH*. Дело в том, что его транскрипты образуют шпильку, которая включает в себя сайт разрезания и DSE. Однако эта вторичная структура не снижает эффективность сайта полиаденилирования *bGH*, как этого можно было бы ожидать на основе нашего исследования и данных, полученных при исследовании терминатора SV40 late [186]. Напротив, было показано, что вторичная структура терминатора *bGH* каким-то образом способствует его активности [528]. Интересно отметить, что если приравнять к нулю значения МСЭ во входных данных для DoSIA, её производительность на данных Goodwin and Rottman, 1992 [503] существенно возрастает, но всё ещё заметно уступает APARENT2. Это не удивительно, так как последовательность терминатора гена *bGH* присутствовала в обучающей выборке APARENT2 [5]. Интересно отметить, что при занулении МСЭ производительность DoSIA на данных Wu and Alwine, 2004 [186] заметно падает, что подтверждает наше предположение о том, что способность DoSIA напрямую учитывать стабильность вторичной структуры в транскриптах терминатора SV40 late обеспечивает её преимущество перед APARENT2 на этих данных.

Далее мы проверили применимость DoSIA для анализа влияния нуклеотидных замен в области DSE на уровень экспрессии генов в различных тканях человека. Для этого были выбраны гены человека, в которых, согласно базе PolyASite [529], более 90% транскриптов используют один сайт полиаденилирования. Было обнаружено 8650 таких генов, что согласуется с данными о том, что у человека около 30% генов имеют единственный сайт полиаденилирования [242]. Для этих генов мы определили последовательности DSE, расположенные в районе с 3 по 38 н. после сайта полиаденилирования. Затем мы провели поиск ЛКП в базе eQTL Catalogue

[530], расположенных в области DSE и связанных с изменением уровня экспрессии этого гена. В анализ были взяты только ЛКП с апостериорной вероятностью того, что они являются каузальными вариантами, 0.25 и более. Было обнаружено 28 ЛКП в 24 генах. В 74% случаев DoSIA правильно предсказала знак изменения уровня экспрессии. APARENT2 правильно предсказал знак изменения уровня экспрессии в 78% случаев. Затем мы сравнили производительность DoSIA и APARENT2 на данных о влиянии ЛКП в области DSE на уровень экспрессии генов в различных тканях человека. Только в семи тканях человека наблюдалась достоверная (р-значение <0.05) или близкая к достоверной (р-значение <0.1) корреляция Пирсона между предсказаниями хотя бы одной из моделей и измеренным уровнем экспрессии (Таблица 8). Это связано с небольшим количеством обнаруженных ЛКП. Интересно отметить, что DoSIA демонстрирует высокий коэффициент корреляции Пирсона (>0.95) для всех семи тканей, тогда как для APARENT2 коэффициент корреляции Пирсона варьировал от -0.85 до >0.99 . Таким образом, можно предположить, что DoSIA подходит для анализа влияния нуклеотидных замен на уровень экспрессии генов человека с одиночными СПА лучше, чем APARENT2, что неудивительно, так как DoSIA была обучена на данных об уровне экспрессии репортёров с одиночными СПА, тогда как APARENT2 на данных об уровне экспрессии репортёров с двумя последовательно расположенными СПА.

Использование DoSIA для предсказания последовательностей DSE, обеспечивающих изменение уровня экспрессии репортёра

Как было упомянуто ранее, эта работа берёт своё начало с обнаружения делеции цитозина (ΔC) в позиции +32 п.н. ниже СПА sNRP-1, которая приводит к двукратному повышению уровня экспрессии репортёрного гена eGFP в культивируемых эмбриональных стволовых клетках мыши mESCs, мышинных клетках 3T3 и клетках человека HEK293T [7,8]. Мы использовали DoSIA для предсказания влияния всех возможных однонуклеотидных замен, делеций и инсерций в исходной последовательности после СПА sNRP-1 на

Таблица 8. DoSIA и APARENT2 были использованы для предсказания влияния ЛКП в области DSE на уровень экспрессии генов в различных тканях человека. Для APARENT2 приведены две пары столбцов, так как она даёт на выходе 2 типа предсказаний. Эти предсказания очень близки и ранее давали идентичные результаты, поэтому в остальных случаях для APARENT2 приведено только одно значение. ЛКП – локус количественных признаков; APARENT – APA REgression NeT; DoSIA – Downstream Sequence element Iterative Analyzer; ρ – коэффициент корреляции Пирсона.

Ткань	DoSIA		APARENT2 (narrow)		APARENT2	
	ρ	p-значение	ρ	p-значение	ρ	p-значение
Большеберцовый нерв	<u>0.9929</u>	<u>0.076</u>	<u>0.9936</u>	<u>0.072</u>	0.9982	0.038
Мозг (мозжечок)	0.9743	0.026	0.6538	0.346	0.6674	0.333
Мозг (гипоталамус)	0.9994	0.021	-0.846	0.358	-0.803	0.406
Мозг (чёрная субстанция)	<u>0.9674</u>	<u>0.163</u>	0.9997	0.015	0.997	0.049
Простата	<u>0.9949</u>	<u>0.064</u>	<u>0.9952</u>	<u>0.063</u>	<u>0.9961</u>	<u>0.056</u>
Сигмовидная ободочная кишка	0.9995	0.02	-0.16	0.898	-0.087	0.945
Яичник	0.952	0.198	0.9983	0.036	<u>0.9939</u>	<u>0.07</u>

уровень экспрессии репортёрного гена eGFP (Рисунок 23, Таблица 9). 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). Мутации +30_T_G не было в обучающей выборке DoSIA.

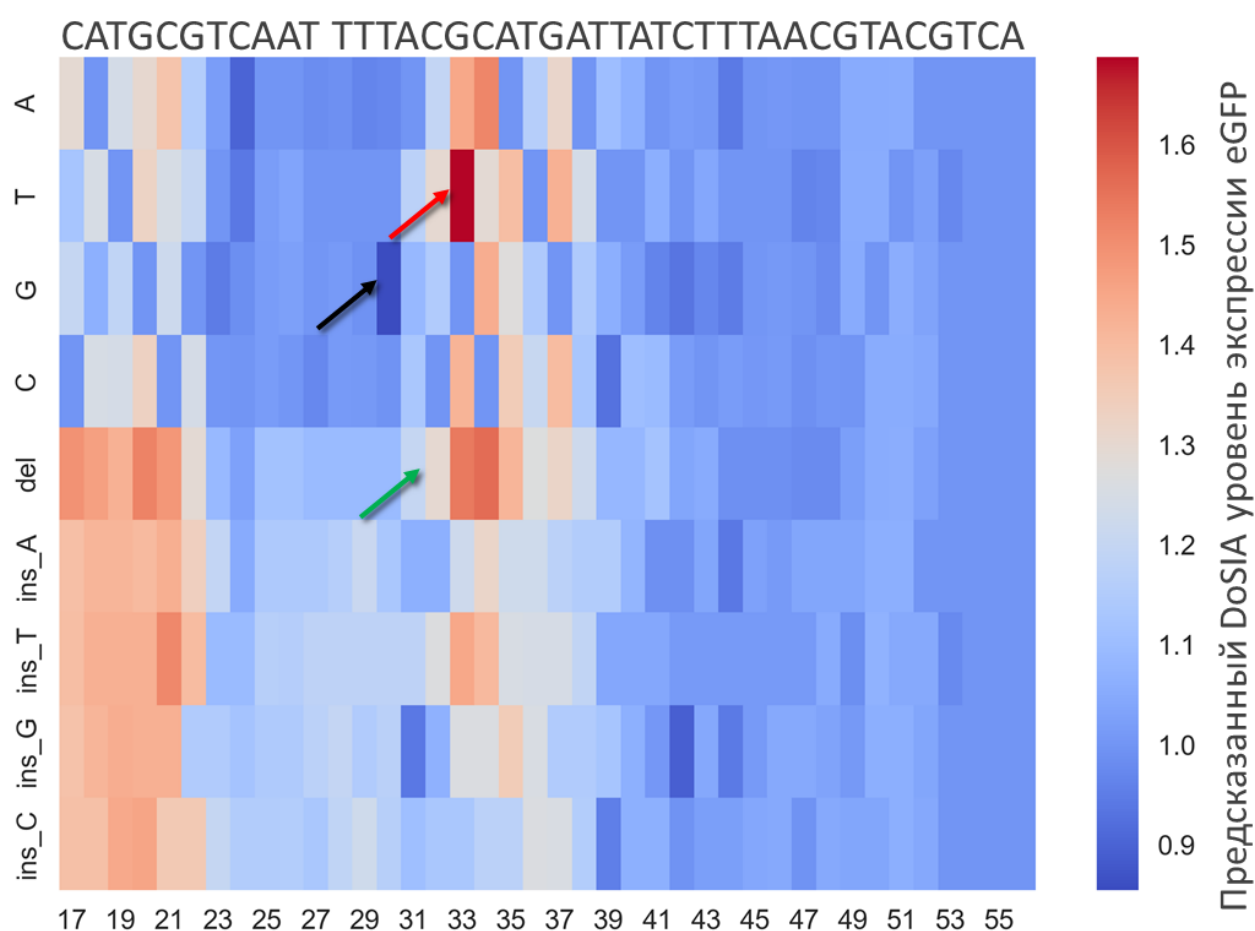


Рисунок 23. Предсказанные DoSIA влияния всех возможных однонуклеотидных замен, делеций и инсерций в исходной последовательности после СПА sNRP-1 (WT) на уровень экспрессии репортёрного гена eGFP. Красная стрелка указывает на мутацию +33_G_T, чёрная – на мутацию +30_T_G, зелёная – на мутацию +31_AC_A (ΔC). DoSIA – Downstream Sequence element Iterative Analyzer; eGFP – улучшенный зелёный флуоресцентный белок; sNRP-1 – растворимый нейропилин 1.

Таблица 9. Предсказанное и наблюдаемое влияние разнообразных замен в области DSE на уровень зрелой мРНК и белка репортёрного гена eGFP. Уровень зрелой мРНК eGFP был измерен методом ОТ-кПЦР, белка – методом проточной цитофлуорометрии. При использовании DoSIA для предсказания уровня экспрессии bGH и его вариантов МСЭ приравнивали к 0. APARENT – APA REgression NeT; bGH – бычий гормон роста; DoSIA – Downstream Sequence element Iterative Analyzer.

Терминатор	Уровень зрелой мРНК eGFP в клетках HEK293T	Уровень белка eGFP в клетках HEK293T	Предсказания DoSIA	Предсказания APARENT2
WT	1	1	1	1
+31_AC_A (ΔC)	1.7	1.9	1.3	1.1
+30_T_G	0.8	0.7	0.9	0.9
+33_G_T	3.3	3.8	1.7	1.3
+29..36_GTCTCTCT	2.9	3	2	1.6
+29..36_GTGTACTT	2.9	3.2	2.2	1.9
+29..36_TTTTCACT	3.2	3.7	2.1	1.9
High1	4.7	4.7	3.9	6.3
High2	5.2	5.4	3.9	6.3
Low1	0.7	0.4	0.7	0.2
SV40 early	3.6	3.1	2.1	4.1
SV40 early +33_T_A	3.5	3.5	1.7	2.5
SV40 early +35_C_CT	4.2	4	2.4	5
bGH	4.7	4.7	2.4	2.8
bGH +25_T_TA	3.8	3.1	2.1	2
bGH +30_A_T	4.1	4.4	2.8	3.8

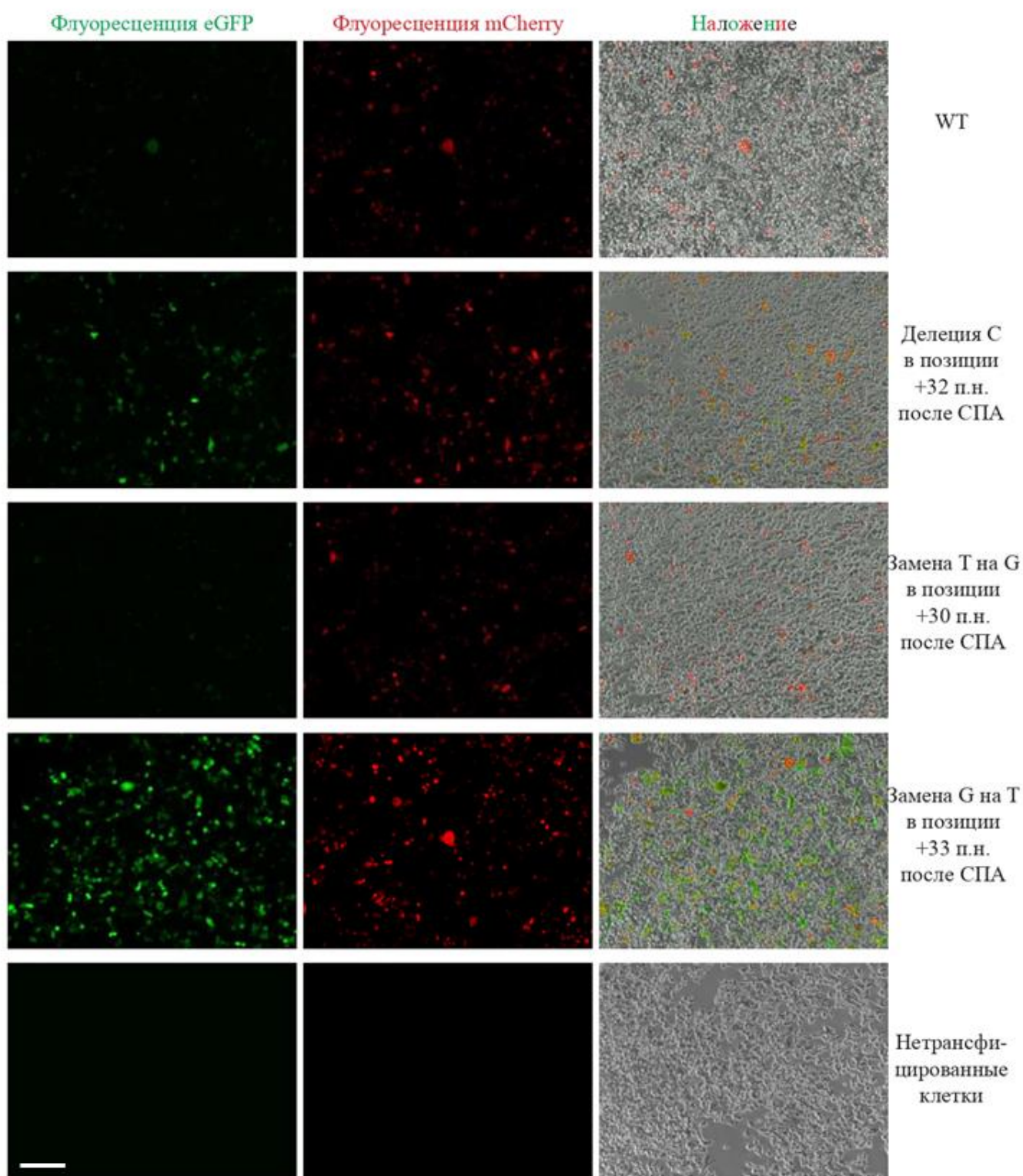


Рисунок 24. Однуклеотидные замены в последовательности WT после СПА оказывают влияние на уровень экспрессии eGFP. Клетки HEK293T были трансфицированы плазмидой рТТС-Hsar-WT (WT) и её вариантами, имеющими однуклеотидные замены после СПА. Фотографии были сделаны через 48 часов после трансфекции. СПА - сигнал полиаденилирования. Масштаб 100 мкм.

Мы получили две плазмидные конструкции с этими нуклеотидными заменами ниже СПА sNRP-1 eGFP. Каждой из полученных конструкций были трансфицированы культивируемые клетки НЕК293Т (Рисунок 24). Через двое суток после трансфекции мы измерили уровень зрелой мРНК eGFP методом ОТ-кПЦР (Рисунок 25, Таблица 9), а уровень белка – методом проточной цитофлуорометрии (Рисунок 26, Таблица 9).

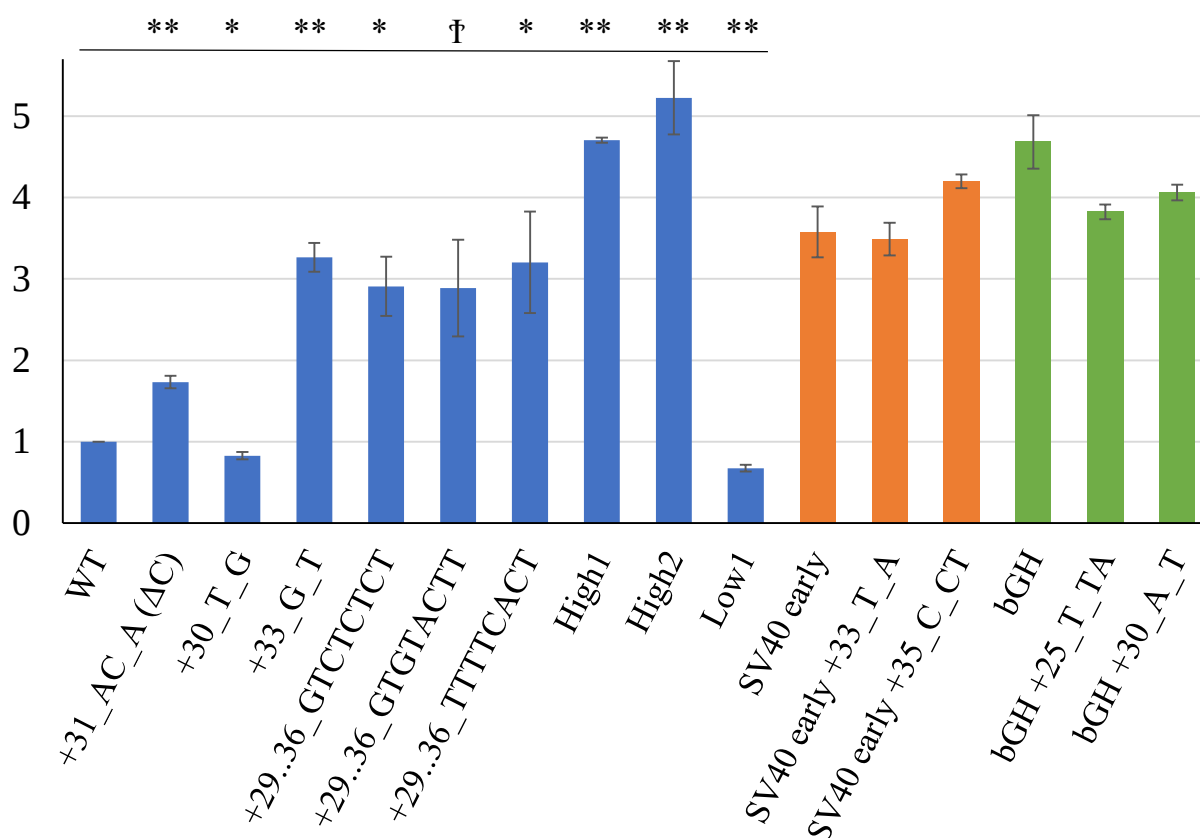


Рисунок 25. Уровень зрелой мРНК eGFP для индивидуальных мутаций, измеренный методом ОТ-кПЦР. Синим изображены значения для последовательности WT и её вариантов, оранжевым – для терминатора транскрипции SV40 early и его вариантов, зелёным – для терминатора транскрипции bGH и его вариантов. Эксперимент был проведён дважды. Пределы погрешности отражают стандартную ошибку. Уровень достоверности был определён с помощью t-критерия Стьюдента при сравнении с конструкцией WT. bGH – бычий гормон роста; eGFP – улучшенный зелёный флуоресцентный белок; sNRP-1 – растворимый нейропилин 1; WT – исходная последовательности после СПА sNRP-1. † – р-значение < 0.1, * – р-значение < 0.05, ** – р-значение < 0.01, *** – р-значение < 0.001.

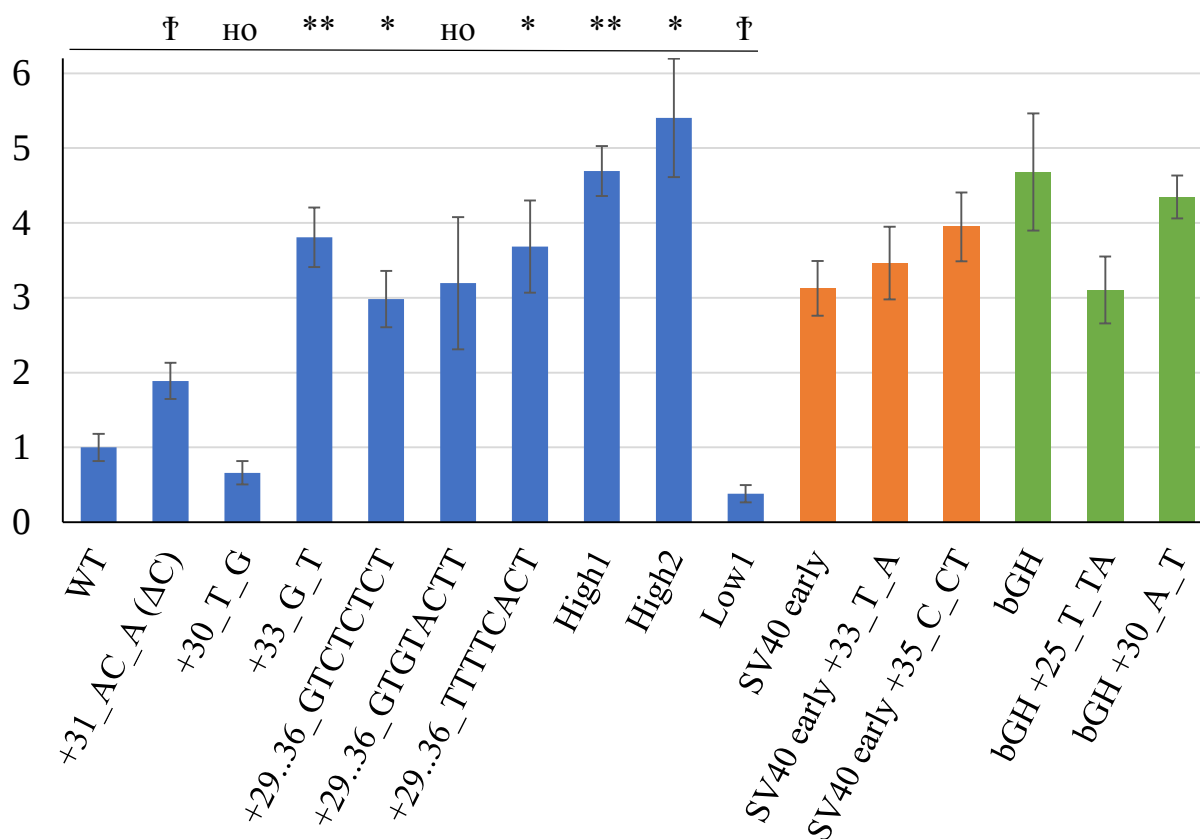


Рисунок 26. Уровень белка eGFP для индивидуальных мутаций, измеренный методом проточной цитофлуорометрии. Синим изображены значения для последовательности WT и её вариантов, оранжевым – для терминатора транскрипции SV40 early и его вариантов, зелёным – для терминатора транскрипции bGH и его вариантов. Эксперимент был проведён трижды. Пределы погрешности отражают стандартную ошибку. Уровень достоверности был определён с помощью t-критерия Стьюдента при сравнении с конструкцией WT. СПА – сигнал полиаденилирования; bGH – бычий гормон роста; eGFP – улучшенный зелёный флуоресцентный белок; sNRP-1 – растворимый нейропилин 1; WT – исходная последовательности после СПА sNRP-1. но – нет отличий (p-значение ≥ 0.1), † – p-значение < 0.1 , * – p-значение < 0.05 , ** – p-значение < 0.01 , *** – p-значение < 0.001 .

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 (р-значение = 0.85, t-критерий Стьюдента) и белка (р-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 Т-богатости области DSE (Рисунок 21), так и с их влиянием на стабильность вторичной структуры пре-мРНК после СПА (Рисунок 22).

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 (Таблица 9). Плазмидами, несущими эти мутации, были трансфицированы клетки HEK293T (Рисунок 27). 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 25) и белка eGFP (Рисунок 26). 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 Т-богатыми. При этом GT- и CT-богатые мотивы в большей степени обогащены в районе +17..+30 н. после СПА, тогда как Т-богатые – в районе +31..+49 н. после СПА sNRP-1, что согласуется с

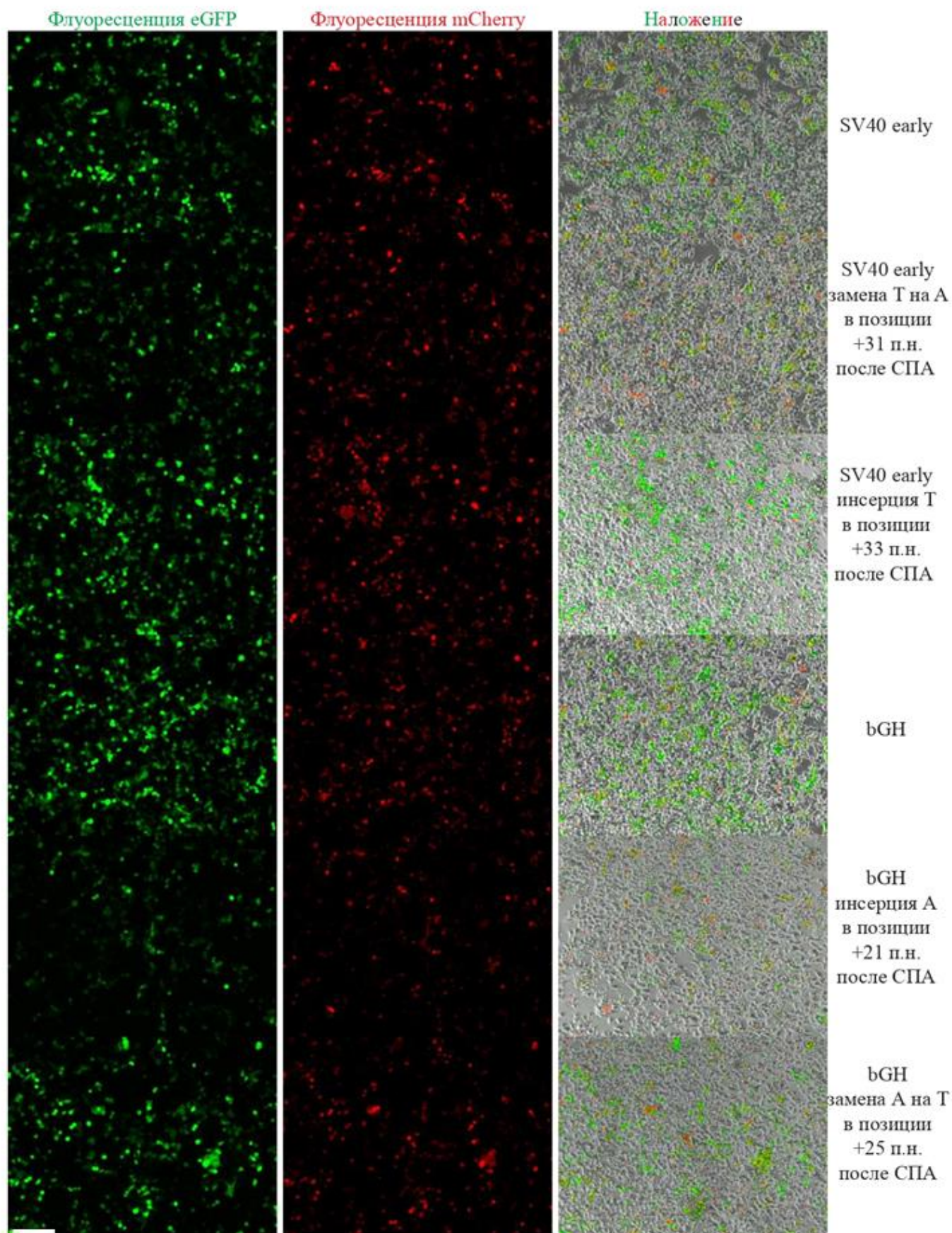


Рисунок 27. Однуклеотидные замены в последовательностях после СПА терминаторов SV40 early и bGH оказывают незначительное влияние на уровень экспрессии eGFP. Клетки HEK293T были трансфицированы плазмидами pTTC-Hsap-SV40 (SV40 early), pTTC-Hsap-bGH (bGH) или их вариантами, имеющими однуклеотидные замены после СПА. Фотографии были сделаны через 48 часов после трансфекции. СПА - сигнал полиаденилирования. Масштаб 100 мкм.

полученными ранее данными о строении DSE в геноме человека [165,172]. Последовательности High1 и High2 обеспечивают низкую стабильность вторичных структур пре-мРНК после СПА (Рисунок 22). Последовательность Low1 является Т-бедной, GC-богатой и обеспечивает формирование стабильной вторичной структуры в пре-мРНК после СПА (Рисунок 22). Последовательности High1 и Low1 не встречаются среди последовательностей, доступных в базе данных «Nucleotide collection» NCBI (<https://www.ncbi.nlm.nih.gov/>). Фрагмент High2 длиной 28 нуклеотидов встречается на шестой хромосоме серебристой сайды (*Pollachius pollachius*). Геном серебристой сайды не аннотирован, однако мы предполагаем, что обнаруженный фрагмент High2 не выполняет функцию DSE, так как мы не обнаружили к в его окружении известные функциональные последовательности СПА, расположенные на той же цепи ДНК. Некоторые восьминуклеотидные участки последовательностей High1, High2 и Low1, расположенные в исследованных районах после СПА sNRP-1, встречаются в соответствующих плазмидных библиотеках (Таблица 10). Восьминуклеотидные участки последовательностей High1, High2 и Low1 встречаются в плазмидных библиотеках, в которых было проанализировано более 25 тысяч мутаций (Таблица 2), то есть более 38% всех возможных последовательностей длиной 8 п.н.

Плазмидами, несущими эти мутации, были трансфицированы клетки НЕК293Т (Рисунок 28). Было показано, что мутации High1 и High2 обеспечивают повышение уровня зрелой мРНК eGFP в 4.7 и 5.2 раза соответственно (Рисунок 25), а уровень белка eGFP в 4.7 и 5.4 раза соответственно (Рисунок 26). Интересно отметить, что мутации High1 и High2 обеспечивают такой же уровень зрелой мРНК (р-значения = 0.77 и 0.81 соответственно, t-критерий Стьюдента) и белка eGFP (р-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 25).

Таблица 10. Некоторые восьминуклеотидные участки последовательностей High1, High2 и Low1 встречаются в соответствующих плазмидных библиотеках.

Библиотека	High1	High2	Low1
17-24	-	-	-
21-28	-	-	-
25-32	Встречается среди 0.2% мутаций с самым высоким уровнем экспрессии по данным МПРА	Встречается среди 0.2% мутаций с самым высоким уровнем экспрессии по данным МПРА	-
29-36	-	-	-
33-40	Встречается среди 4% мутаций с самым высоким уровнем экспрессии по данным МПРА	-	-
37-44	Встречается среди 0.1% мутаций с самым высоким уровнем экспрессии по данным МПРА	Встречается среди 0.1% мутаций с самым высоким уровнем экспрессии по данным МПРА	-
41-48	Встречается среди 16% мутаций с самым высоким уровнем экспрессии по данным МПРА	Встречается среди 16% мутаций с самым высоким уровнем экспрессии по данным МПРА	Встречается среди 1% мутаций с самым низким уровнем экспрессии по данным МПРА
45-52	Встречается среди 28% мутаций с самым низким уровнем экспрессии по данным МПРА	Встречается среди 28% мутаций с самым низким уровнем экспрессии по данным МПРА	-

Корреляция Пирсона между предсказаниями DoSIA и данными ОТ кПЦР составила 0.9, между предсказаниями APARENT2 и данными ОТ кПЦР – 0.82, между предсказаниями DoSIA и данными проточной цитофлуорометрии – 0.89, между предсказаниями APARENT2 и данными проточной цитофлуорометрии – 0.77 (Таблица 9). Корреляция Пирсона между данными ОТ-кПЦР и проточной цитофлуорометрии составила 0.97. Таким образом, в данной системе уровень зрелой мРНК репортёрного гена определяет уровень образующегося белка, а DoSIA является лучшим инструментом для предсказания влияния последовательности DSE на уровень экспрессии репортёра.

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 и полиаденилирования пре-мРНК. Для того, чтобы проверить это предположение, мы измерили количество непротессированных транскриптов для всех исследованных конструкций. Для этого мы провели обратную транскрипцию с геноспецифического праймера RasI-R (Приложение 2) с последующей кПЦР. Праймер RasI-R комплементарен области, расположенной на расстоянии от +97 до +137 н. после сайта разрезания.

Оказалось, что мутация Low1 обеспечивает более высокий уровень непротессированной мРНК eGFP по сравнению с мутациями High1 и High2 (p-значения = 0.048 и 0.047 соответственно, t-критерий Стьюдента, Рисунок 29). Между уровнями незрелых и полиаденилированных транскриптов для всех исследованных конструкций наблюдается обратная корреляция Пирсона, равная -0.69, то есть чем больше зрелой мРНК eGFP, тем меньше непротессированной (неполиаденилированной) мРНК eGFP. Таким образом, последовательность после сайта полиаденилирования действительно оказывает влияние на эффективность процессинга пре-мРНК, что, в свою очередь, влияет на уровень зрелой мРНК и, в конечном счёте, белка.

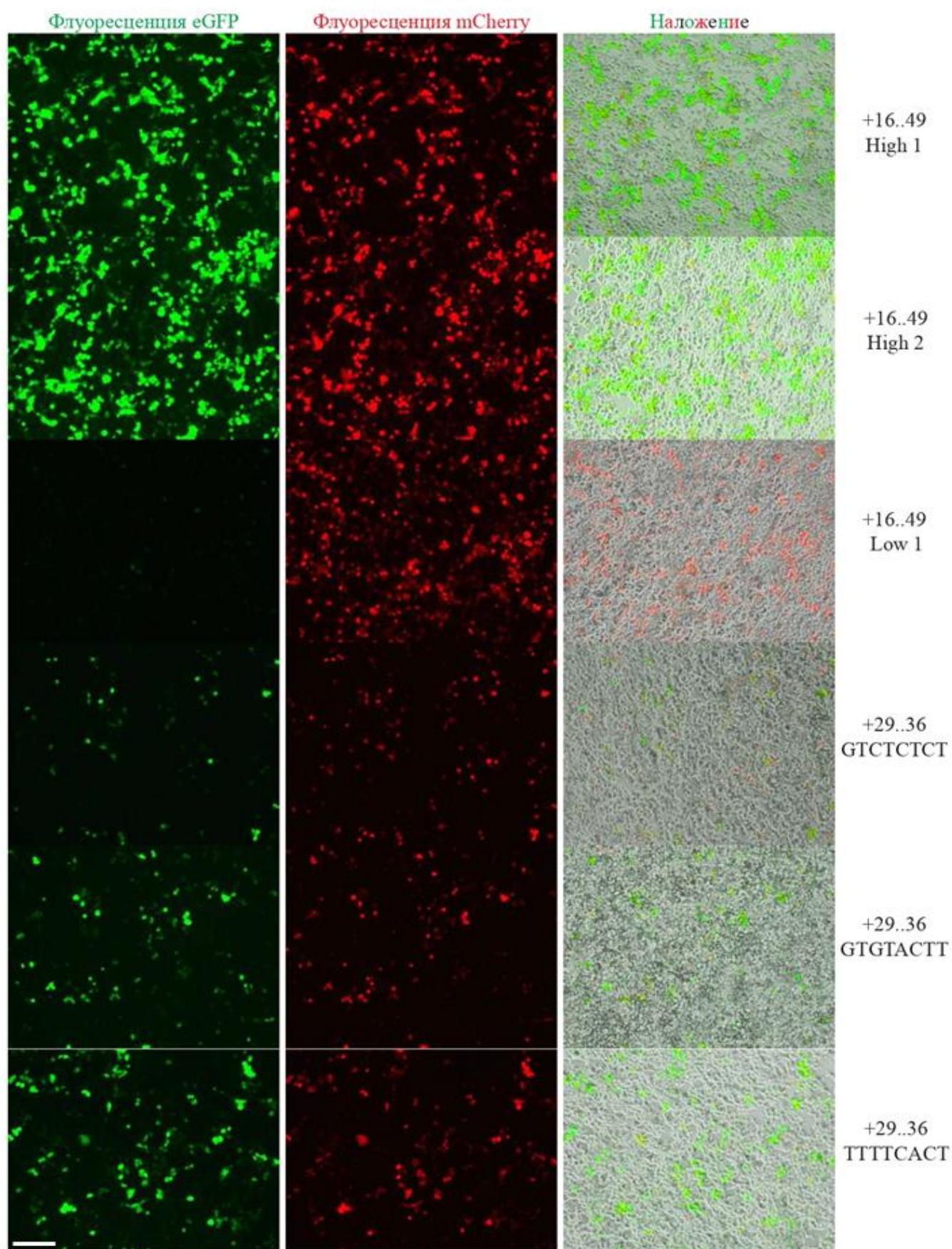


Рисунок 28. При замене последовательности WT в районе +16..49 н. после СПА наблюдалось варьирование уровня экспрессии eGFP до 7 раз. Клетки HEK293T были трансфицированы вариантами плазмиды rTTC-Hsap-WT (WT), имеющими замены после СПА. Фотографии были сделаны через 48 часов после трансфекции. СПА - сигнал полиаденилирования. Масштаб 100 мкм.

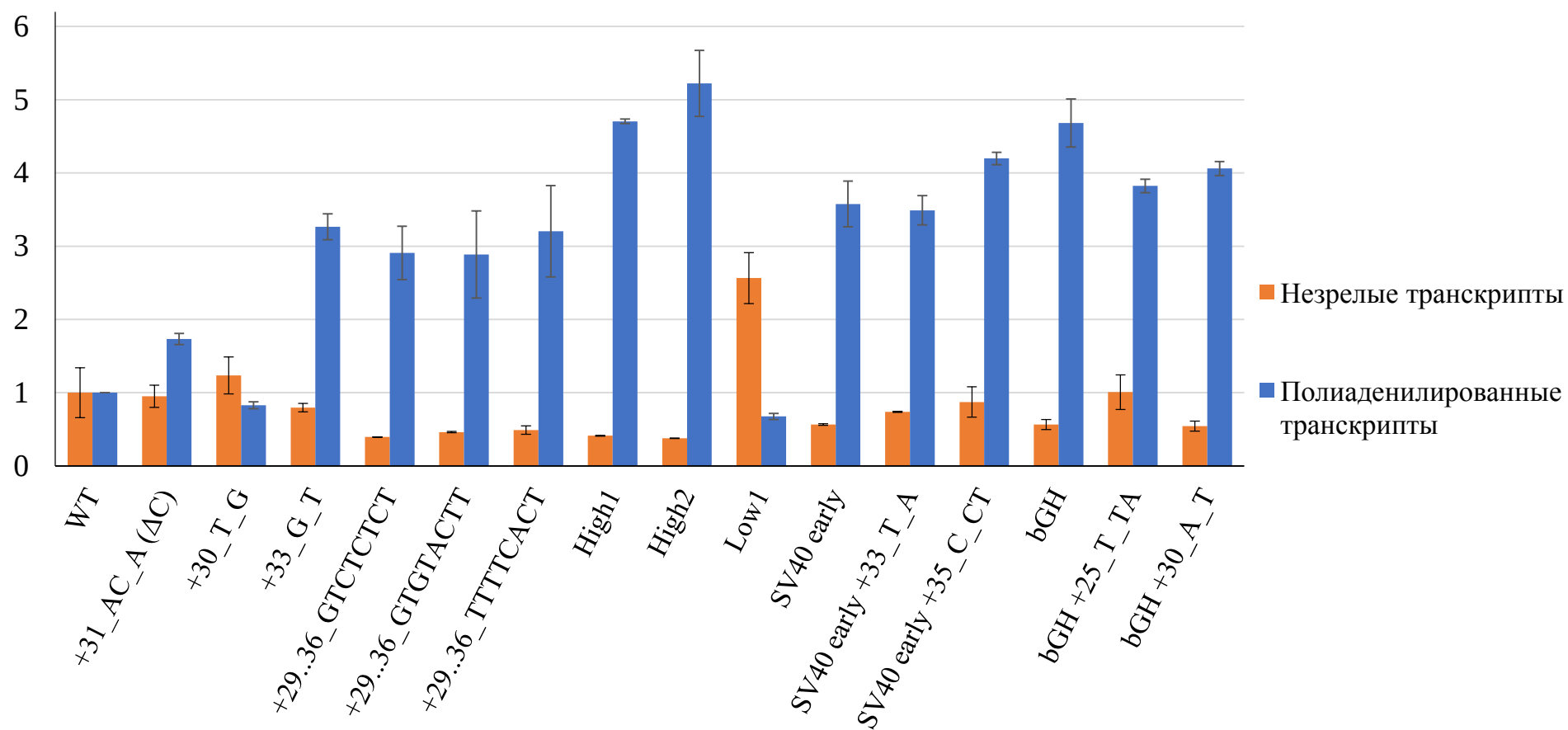


Рисунок 29. Уровень незрелой (непроцессированной) и зрелой (полиаденилированной) мРНК eGFP для индивидуальных мутаций, измеренный методом ОТ-кПЦР. Эксперимент был проведён дважды. Пределы погрешности отражают стандартную ошибку. СПА – сигнал полиаденилирования; bGH – бычий гормон роста; eGFP – улучшенный зелёный флуоресцентный белок; sNRP-1 – растворимый нейропилин 1; WT – исходная последовательности после СПА sNRP-1.

Заключение

Полиаденилирование – важный этап процессинга 3'-конца пре-мРНК эукариот. Сложность белковых комплексов, осуществляющих процессинг 3'-конца пре-мРНК, наводит на мысль о том, что процесс разрезания и полиаденилирования не только обеспечивает формирование стабильной, способной к трансляции мРНК, но и играет важную регуляторную роль, влияя на уровень экспрессии генов. Впервые влияние последовательности терминатора транскрипции на уровень зрелой мРНК и белка репортёрного гена было показано более 15 лет назад в лаборатории небезызвестного Nicholas Proudfoot [2]. Однако регуляторный потенциал терминаторов транскрипции РНК-полимеразы II по-прежнему остаётся недооценённым и непонятым широкой научной общественностью.

В данной работе было исследовано влияние последовательности, расположенной в районе +17..56 п.н. после СПА, на уровень зрелой матричной РНК репортёрного гена eGFP в культивируемых клетках человека HEK293T методом МПРА. Этот район соответствует DSE – цис-элементу терминатора транскрипции, соответствующему сайту связывания белка CstF64 – компоненту канонического комплекса процессинга 3'-конца пре-мРНК [62]. Белок CstF64 входит в состав комплекса CstF, необходимого для разрезания и полиаденилирования 3'-конца пре-мРНК.

В данной работе было показано, что последовательность и вторичная структура РНК в области DSE оказывают большое влияние на уровень зрелой мРНК и белка eGFP. Было показано, что оптимальная последовательность DSE обеспечивает эффективный процессинг 3'-конца пре-мРНК и снижает долю непроцессированных транскриптов, которые могут возникать при преждевременной терминации транскрипции РНК-полимеразы II независимо от разрезания 3'-конца пре-мРНК [297]. Неполиаденилированные

транскрипты не могут транспортироваться в цитоплазму и быть транслированы [67]. Это делает терминаторы транскрипции РНК-полимеразы II важными регуляторными элементами.

В данной работе было показано, что для последовательностей после СПА, связанных с высоким уровнем зрелой мРНК гена eGFP в клетках НЕК293Т, характерна Т-обогащённость и наличие GT- и СТ-богатых мотивов, характерных для сайтов связывания белка CstF64 [126,128,129]. Особый интерес вызывают мотивы, характерные для последовательностей после СПА, связанных с низким уровнем зрелой мРНК репортёрного гена в клетках НЕК293Т. Они GC-богатые, часто содержат мотив AGG. Точно такие же мотивы характерны и для штрихкодов, связанных с низким уровнем экспрессии репортёра. Они напоминают последовательность MAZ GGGGGAGGGGG, способствующую терминации транскрипции РНК-полимеразы II в клетках млекопитающих [288,531,532]. Интересно отметить, что последовательность MAZ может повышать уровень зрелой мРНК и белка репортёра в том случае, если она расположена на расстоянии ~800 п.н. после СПА [2]. Мы предполагаем, что GC-богатые последовательности, содержащие мотив AGG, способствуют преждевременной терминации транскрипции РНК-полимеразы II в клетках млекопитающих, что снижает уровень экспрессии репортёрного гена.

В ходе данной работы была обучена модель DoSIA, позволяющая предсказывать уровень зрелой мРНК репортёрного гена в зависимости от последовательности его DSE. DoSIA имеет ряд ограничений по сравнению с имеющимися моделями, предсказывающими уровень экспрессии гена в зависимости от последовательности терминатора транскрипции, таких как APARENT2. Предсказания DoSIA ограничиваются регуляторным потенциалом только одного цис-элемента – DSE. DoSIA не может учитывать взаимные влияния последовательности DSE и других цис-элементов, вовлечённых в процессинг 3'-конца пре-мРНК. Однако судя по тому, что

предсказательная сила DoSIA сравнима с APARENT2, можно предположить, что эти взаимные влияния не играют большой роли. DoSIA едва ли может быть применима для генов с несколькими альтернативными сайтами полиаденилирования. DoSIA также имеет ряд преимуществ. Она напрямую учитывает влияние вторичной структуры пре-мРНК после СПА на уровень экспрессии гена, что даёт ей преимущество перед APARENT2 на некоторых наборах данных. DoSIA, судя по всему, превосходит APARENT2 при предсказании влияния последовательности DSE на уровень экспрессии генов с одиночными СПА (Таблица 8).

DoSIA позволила предсказать однонуклеотидную замену G на T в исходной последовательности в позиции +33 п.н. после СПА sNRP-1, которая повысила уровень зрелой мРНК eGFP в 3.5 раза до уровня, который обеспечивает широко используемый в научных исследованиях и биотехнологии терминатор SV40 early. С помощью DoSIA также были предсказаны последовательности DSE High1 и High2, повышающие уровень зрелой мРНК и белка eGFP в клетках HEK293T в 4.7 и 5.2 раза соответственно, и последовательность Low1, понижающая уровень зрелой мРНК eGFP в клетках HEK293T в 1.4 раза. Таким образом, последовательность DSE позволяет варьировать уровень зрелой мРНК репортёрного гена в клетках HEK293T как минимум в 7.4 раза.

Сравнение данных проведённого нами МПРА и TRIP (тысячи репортёров, интегрированных параллельно), проведённого со схожими плазмидными библиотеками Akhtar, de Jong, Pindyurin и другими [7], позволяет сделать вывод, что регуляторный потенциал последовательности DSE примерно на один порядок ниже регуляторного потенциала эффекта положения. Однако, за неимением данных об обратном, можно сделать вывод, что влияние последовательности DSE и эффекта положения на уровень экспрессии гена независимо друг от друга. Последовательность репортёрного гена также не оказывает влияния на регуляторный эффект DSE [519].

Последовательность промотора может влиять на ширину динамического диапазона влияния DSE на уровень экспрессии гена так же, как это было показано для элементов, паузирующих РНК-полимеразу II [2]. При этом для промоторов, обеспечивающих высокую скорость элонгации, таких как промотор вируса иммунодефицита человека, особенно важно сочетание с высокоэффективными сигналами полиаденилирования и элементами, паузирующими РНК-полимеразу II. Напротив, промоторы, обеспечивающие относительно низкую скорость элонгации, такие как промотор hCMV, могут обеспечивать высокий уровень экспрессии репортёрного гена даже в сочетании с неоптимальными сигналами полиаденилирования.

DoSIA может быть использована как инструмент для разработки новых синтетических терминаторов транскрипции РНК-полимеразы II, обеспечивающих желаемый уровень зрелой мРНК и белка целевого гена. Исследованный в данной работе регуляторный район очень компактный (всего 36 п.н.) и расположен между сайтами рестрикции SpeI и BsiWI в исходной последовательности после СПА sNRP-1 (Приложение 1), что делает клонирование новых синтетических терминаторов транскрипции очень удобным.

Однако, несмотря на стремительное развитие моделей, позволяющих разрабатывать новые высокоэффективные терминаторы транскрипции, исследователи и биотехнологи не торопятся применять их в своей работе. Наиболее популярным на сегодняшний день синтетическим терминатором транскрипции является терминатор, разработанный ещё в конце 80-х в лаборатории Nicholas Proudfoot [533]. Он привлекает исследователей благодаря своей высокой эффективности и небольшой длине (37 п.н.) [3,534]. Он получен путём встройки между СПА и DSE терминатора транскрипции β -глобина кролика двух сайтов элонгуазы рестрикции BglII, разделённых между собой Т-богатой последовательностью. Иронично, что этот терминатор был разработан в исследовательских целях и его высокая эффективность

оказалась большой неожиданностью для его создателей. В остальном большинство исследователей предпочитают использовать терминаторы транскрипции генов с высоким уровнем экспрессии, такие как терминаторы SV40 early [3], SV40 late, генов bGH [3], β -глобина человека или кролика, либо комбинировать их части [450].

Сегодня мы имеем на руках все карты для разработки синтетических терминаторов транскрипции РНК-полимеразы II, позволяющих модулировать эффективность процессинга 3'-конца пре-мРНК (DoSIA, APARENT2), выбор сайта полиаденилирования (APARENT) и соотношение продуктов АПА (APARENT, PolyStrength). И то, что Nicholas Proudfoot, несмотря на свой первоначальный скепсис в отношении биоинформатических подходов к исследованию процессинга 3'-конца пре-мРНК и терминации транскрипции, использовал APARENT для анализа экспериментальных данных в своей недавней работе [535], вселяет надежду, что рано или поздно упомянутые модели, в том числе обученная нами DoSIA, будут использованы для разработки обладающих желаемыми свойствами синтетических терминаторов транскрипции РНК-полимеразы II.

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, связанных с высоким уровнем зрелой мРНК гена eGFP в культивируемых клетках человека HEK293T, характерна Т-обогащённость и наличие мотивов CTC, TCT, TGT, TTG, GTCT, GTGT, TGTC, TCTC. Для последовательностей, 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.

Благодарности

Я хочу сказать спасибо всем своим учителям, которые часто меня вдохновляли и во многом предопределили путь, который привёл меня к написанию этой работы.

Я благодарю бывшего с.н.с. лаборатории иммуногенетики ИМКБ СО РАН Горчакова А.А. за помощь в поиске лаборатории. Я благодарю с.н.с. лаборатории эпигенетики НГУ Лактионова П.П. и н.с. лаборатории геномики ИМКБ СО РАН Максимова Д.А. за помощь в запуске Illumina MiSeq. Я благодарю свою первую научную руководительницу, заведующую лабораторией моделирования патологий человека ИМКБ СО РАН Кожевникову Е.Н., под руководством которой я освоила большую часть методов, которыми я владею на сегодняшний день. Я благодарю своего второго научного руководителя, предыдущего заведующего лабораторией клеточного деления ИМКБ СО РАН Пиндюрина А.В. за то, что пригласил меня в этот проект и обучил методу МПРА. Я благодарю научного руководителя этой работы, заведующую лабораторией клеточного деления ИМКБ СО РАН Омелину Е.С. за безграничную помощь и большую творческую свободу. Я благодарю всех своих коллег из лаборатории клеточного деления ИМКБ СО РАН, нынешних и предыдущих, за множество ценных советов и готовность помочь. Мне хочется поблагодарить отдельно заведующую лабораторией нейрогенетики когнитивных и аффективных расстройств НИИНМ Болдыреву Л.В., н.с. лаборатории клеточного деления ИМКБ СО РАН Яринич Л.А. и с.н.с. лаборатории клеточного деления ИМКБ СО РАН Андрееву Е.Н.

Последними, но не по важности, я благодарю членов своей семьи: маму, папу, Славу, Глеба и Лизу, за безграничную помощь, поддержку, и вечный вопрос «Когда же ты, наконец, защитишься?».

Список литературы

1. Proudfoot NJ, Brownlee GG. 3' non-coding region sequences in eukaryotic messenger RNA // *Nature*. 1976. Vol. 263. P. 211–214.
2. West S., Proudfoot N.J. Transcriptional Termination Enhances Protein Expression in Human Cells // *Mol. Cell*. Elsevier Inc., 2009. Vol. 33, № 3. P. 354–364.
3. Wang X.Y. et al. Enhanced Transgene Expression by Optimization of Poly A in Transfected CHO Cells // *Front. Bioeng. Biotechnol.* 2022. Vol. 10, № January. P. 1–9.
4. Zhou L.Y. et al. Optimization of mammalian expression vector by cis-regulatory element combinations // *Mol. Genet. Genomics*. Springer Berlin Heidelberg, 2023. Vol. 298, № 5. P. 1121–1133.
5. Bogard N. et al. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation // *Cell*. 2019. Vol. 178, № 1. P. 91-106.e23.
6. Slutskin I.V., Weinberger A., Segal E. Sequence determinants of polyadenylation-mediated regulation // *Genome Res*. 2019. Vol. 29, № 10. P. 1635–1647.
7. Akhtar W. et al. Chromatin position effects assayed by thousands of reporters integrated in parallel // *Cell*. Elsevier Inc., 2013. Vol. 154, № 4. P. 914–927.
8. Boldyreva L. V. et al. Fine gene expression regulation by minor sequence variations downstream of the polyadenylation signal // *Mol. Biol. Rep.* 2021. Vol. 48, № 2. P. 1539–1547.
9. Greger I.H. et al. Transcriptional interference perturbs the binding of Sp1 to the HIV-1 promoter // *Nucleic Acids Res.* 1998. Vol. 26, № 5. P. 1294–1300.
10. Mapendano C.K. et al. Crosstalk between mRNA 3' End Processing and Transcription Initiation // *Mol. Cell*. 2010. Vol. 40, № 3. P. 410–422.
11. Camblong J. et al. Antisense RNA Stabilization Induces Transcriptional Gene Silencing via Histone Deacetylation in *S. cerevisiae* // *Cell*. 2007. Vol. 131, № 4. P.

706–717.

12. Gullerova M., Proudfoot N.J. Cohesin Complex Promotes Transcriptional Termination between Convergent Genes in *S. pombe* // *Cell*. 2008. Vol. 132, № 6. P. 983–995.
13. Porrua O., Libri D. Transcription termination and the control of the transcriptome: Why, where and how to stop // *Nat. Rev. Mol. Cell Biol.* Nature Publishing Group, 2015. Vol. 16, № 3. P. 190–202.
14. Lesnik E.A., Freier S.M. Relative Thermodynamic Stability of DNA, RNA, and DNA:RNA Hybrid Duplexes: Relationship with Base Composition and Structure // *Biochemistry*. 1995. Vol. 34, № 34. P. 10807–10815.
15. Martin F.H., Tinoco I. DNA-RNA hybrid duplexes containing oligo(dA:rU) sequences are exceptionally unstable and may facilitate termination of transcription // *Nucleic Acids Res.* 1980. Vol. 8, № 10.
16. Peters J.M., Vangeloff A.D., Landick R. Bacterial transcription terminators: The RNA 3'-end chronicles // *Journal of Molecular Biology*. 2011. Vol. 412, № 5.
17. Cardinale C.J. et al. Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in *E. coli* // *Science* (80-.). 2008. Vol. 320, № 5878.
18. Adhya S., Gottesman M. Control of transcription termination. // *Annual review of biochemistry*. 1978. Vol. 47. P. 967–996.
19. Peters J.M. et al. Rho and NusG suppress pervasive antisense transcription in *Escherichia coli* // *Genes Dev.* 2012. Vol. 26, № 23.
20. Allfano P. et al. A consensus motif common to all rho-dependent prokaryotic transcription terminators // *Cell*. 1991. Vol. 64, № 3.
21. Washburn R.S., Gottesman M.E. Regulation of transcription elongation and termination // *Biomolecules*. 2015. Vol. 5, № 2.
22. Porrua O., Boudvillain M., Libri D. Transcription Termination: Variations on Common Themes // *Trends Genet.* Elsevier Ltd, 2016. Vol. 32, № 8. P. 508–522.
23. Boudvillain M., Figueroa-Bossi N., Bossi L. Terminator still moving forward:

- Expanding roles for Rho factor // *Current Opinion in Microbiology*. 2013. Vol. 16, № 2.
24. Figueroa-Bossi N. et al. RNA remodeling by bacterial global regulator CsrA promotes Rho-dependent transcription termination // *Genes Dev*. 2014. Vol. 28, № 11.
 25. Proshkin S., Mironov A., Nudler E. Riboswitches in regulation of rho-dependent transcription termination // *Biochim. Biophys. Acta - Gene Regul. Mech*. 2014. Vol. 1839, № 10. P. 974–977.
 26. Matsumoto Y. et al. Autogenous regulation of the gene for transcription termination factor rho in *Escherichia coli*: Localization and function of its attenuators // *J. Bacteriol*. 1986. Vol. 166, № 3.
 27. Peters J.M. et al. Rho directs widespread termination of intragenic and stable RNA transcription // *Proc. Natl. Acad. Sci. U. S. A*. 2009. Vol. 106, № 36.
 28. Nicolas P. et al. Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis* // *Science* (80-.). 2012. Vol. 335, № 6072.
 29. Leela J.K. et al. Rho-dependent transcription termination is essential to prevent excessive genome-wide R-loops in *Escherichia coli* // *Proc. Natl. Acad. Sci. U. S. A*. 2013. Vol. 110, № 1.
 30. Brennan C.A., Dombroski A.J., Platt T. Transcription termination factor rho is an RNA-DNA helicase // *Cell*. 1987. Vol. 48, № 6.
 31. Washburn R.S., Gottesman M.E. Transcription termination maintains chromosome integrity // *Proc. Natl. Acad. Sci. U. S. A*. 2011. Vol. 108, № 2.
 32. Epshtein V. et al. An allosteric mechanism of Rho-dependent transcription termination // *Nature*. 2010. Vol. 463, № 7278.
 33. Proudfoot N.J. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut // *Science* (80-.). 2016. Vol. 352, № 6291.
 34. Németh A. et al. RNA polymerase I termination: Where is the end? // *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. 2013. Vol. 1829, № 3–4.

35. Reiter A. et al. The Reb1-homologue Ydr026c/Nsi1 is required for efficient RNA polymerase I termination in yeast // *EMBO J.* 2012. Vol. 31, № 16.
36. Merkl P. et al. Binding of the Termination Factor Nsi1 to Its Cognate DNA Site Is Sufficient To Terminate RNA Polymerase I Transcription In Vitro and To Induce Termination In Vivo // *Mol. Cell. Biol.* 2014. Vol. 34, № 20.
37. Mason S.W., Wallisch M., Grummt I. RNA polymerase I transcription termination: Similar mechanisms are employed by yeast and mammals // *Journal of Molecular Biology.* 1997. Vol. 268, № 2.
38. Tschochne H., Milkereit P. RNA polymerase I from *S. cerevisiae* depends on an additional factor to release terminated transcripts from the template // *FEBS Lett.* 1997. Vol. 410, № 2–3.
39. Prescott E.M. et al. Transcriptional termination by RNA polymerase I requires the small subunit Rpa12p // *Proc. Natl. Acad. Sci. U. S. A.* 2004. Vol. 101, № 16.
40. Zhao J., Hyman L., Moore C. Formation of mRNA 3' Ends in Eukaryotes: Mechanism, Regulation, and Interrelationships with Other Steps in mRNA Synthesis // *Microbiol. Mol. Biol. Rev.* 1999. Vol. 63, № 2.
41. Mandel C.R., Bai Y., Tong L. Protein factors in pre-mRNA 3'-end processing // *Cellular and Molecular Life Sciences.* 2008. Vol. 65, № 7–8.
42. Xiang K., Tong L., Manley J.L. Delineating the Structural Blueprint of the Pre-mRNA 3'-End Processing Machinery // *Mol. Cell. Biol.* 2014. Vol. 34, № 11. P. 1894–1910.
43. Shi Y., Manley J.L. The end of the message: Multiple protein–RNA interactions define the mRNA polyadenylation site // *Genes Dev.* 2015. Vol. 29, № 9.
44. Neve J. et al. Cleavage and polyadenylation: Ending the message expands gene regulation // *RNA Biol.* 2017. Vol. 14, № 7. P. 865–890.
45. Tudek A., Lloret-Llinares M., Heick Jensen T. The multitasking polyA tail: Nuclear RNA maturation, degradation and export // *Philosophical Transactions of the Royal Society B: Biological Sciences.* 2018. Vol. 373, № 1762.

46. Stewart M. Polyadenylation and nuclear export of mRNAs // J. Biol. Chem. 2019. Vol. 294, № 9.
47. Cramer P., Bushnell D.A., Kornberg R.D. Structural basis of transcription: RNA polymerase II at 2.8 ångstrom resolution // Science (80-.). 2001. Vol. 292, № 5523.
48. Bartkowiak B., MacKellar A.L., Greenleaf A.L. Updating the CTD Story: From Tail to Epic // Genet. Res. Int. 2011. Vol. 2011.
49. Hsin J.P., Manley J.L. The RNA polymerase II CTD coordinates transcription and RNA processing // Genes and Development. 2012. Vol. 26, № 19.
50. Egloff S., Dienstbier M., Murphy S. Updating the RNA polymerase CTD code: Adding gene-specific layers // Trends in Genetics. 2012. Vol. 28, № 7.
51. Zhang D.W. et al. Emerging Views on the CTD Code // Genet. Res. Int. 2012. Vol. 2012.
52. Heidemann M. et al. Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription // Biochimica et Biophysica Acta - Gene Regulatory Mechanisms. 2013. Vol. 1829, № 1.
53. Buratowski S. Progression through the RNA Polymerase II CTD Cycle // Molecular Cell. 2009. Vol. 36, № 4.
54. Tietjen J.R. et al. Chemical-genomic dissection of the CTD code // Nat. Struct. Mol. Biol. 2010. Vol. 17, № 9.
55. Mayer A. et al. Uniform transitions of the general RNA polymerase II transcription complex // Nat. Struct. Mol. Biol. 2010. Vol. 17, № 10.
56. Kim H. et al. Gene-specific RNA polymerase II phosphorylation and the CTD code // Nat. Struct. Mol. Biol. 2010. Vol. 17, № 10.
57. Bataille A.R. et al. A Universal RNA Polymerase II CTD Cycle Is Orchestrated by Complex Interplays between Kinase, Phosphatase, and Isomerase Enzymes along Genes // Mol. Cell. 2012. Vol. 45, № 2.
58. Mayer A. et al. CTD Tyrosine phosphorylation impairs termination factor

- recruitment to RNA polymerase II // *Science* (80-.). 2012. Vol. 336, № 6089.
59. Rodríguez-Molina J.B., West S., Passmore L.A. Knowing when to stop: Transcription termination on protein-coding genes by eukaryotic RNAPII // *Mol. Cell*. 2023. Vol. 83, № 3. P. 404–415.
 60. Shi Y. et al. Molecular Architecture of the Human Pre-mRNA 3' Processing Complex // *Mol. Cell*. 2009. Vol. 33, № 3.
 61. Proudfoot N.J. Ending the message: Poly(A) signals then and now // *Genes Dev*. 2011. Vol. 25, № 17. P. 1770–1782.
 62. Sun Y., Hamilton K., Tong L. Recent molecular insights into canonical pre-mRNA 3'-end processing // *Transcription*. Taylor & Francis, 2020. Vol. 11, № 2. P. 83–96.
 63. Chan S.L. et al. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing // *Genes Dev*. 2014. Vol. 28, № 21.
 64. Schönemann L. et al. Reconstitution of CPSF active in polyadenylation: Recognition of the polyadenylation signal by WDR33 // *Genes Dev*. 2014. Vol. 28, № 21.
 65. MacDonald C.C., Wilusz J., Shenk T. The 64-Kilodalton Subunit of the CstF Polyadenylation Factor Binds to Pre-mRNAs Downstream of the Cleavage Site and Influences Cleavage Site Location // *Mol. Cell. Biol*. 1994. Vol. 14, № 10. P. 6647–6654.
 66. Rügsegger U., Beyer K., Keller W. Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors // *J. Biol. Chem*. 1996. Vol. 271, № 11.
 67. Liu J. et al. Molecular Insights into mRNA Polyadenylation and Deadenylation // *Int. J. Mol. Sci*. 2022. Vol. 23, № 19.
 68. Boreikaitė V., Passmore L.A. 3'-End Processing of Eukaryotic mRNA: Machinery, Regulation, and Impact on Gene Expression // *Annu. Rev. Biochem*. 2023. Vol. 92. P. 199–225.
 69. Venkataraman K., Brown K.M., Gilmartin G.M. Analysis of a noncanonical

- poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition // *Genes Dev.* 2005. Vol. 19, № 11.
70. Nunes N.M. et al. A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence // *EMBO J.* 2010. Vol. 29, № 9. P. 1523–1536.
71. Higgs D.R. et al. α -Thalassaemia caused by a polyadenylation signal mutation // *Nature.* 1983. Vol. 306, № 5941.
72. Orkin S.H. et al. Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. // *EMBO J.* 1985. Vol. 4, № 2.
73. Bennett C.L. et al. A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA→AAUGAA) leads to the IPEX syndrome // *Immunogenetics.* 2001. Vol. 53, № 6.
74. Whitelaw E., Proudfoot N. Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene. // *EMBO J.* 1986. Vol. 5, № 11.
75. Stacey S.N. et al. A germline variant in the TP53 polyadenylation signal confers cancer susceptibility // *Nat. Genet.* 2011. Vol. 43, № 11.
76. Garin I. et al. Recessive mutations in the INS gene result in neonatal diabetes through reduced insulin biosynthesis // *Proc. Natl. Acad. Sci. U. S. A.* 2010. Vol. 107, № 7.
77. Wiestner A. et al. Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival // *Blood.* 2007. Vol. 109, № 11.
78. Russo P. et al. Signals that produce 3' termini in CYC1 mRNA of the yeast *Saccharomyces cerevisiae*. // *Mol. Cell. Biol.* 1993. Vol. 13, № 12.
79. Graber J.H. et al. In silico detection of control signals: MRNA 3'-end-processing sequences in diverse species // *Proc. Natl. Acad. Sci. U. S. A.* 1999. Vol. 96, № 24.
80. Dantonel J.C. et al. Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA // *Nature.* 1997. Vol. 389, № 6649. P. 399–402.

81. Glover-Cutter K. et al. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes // *Nat. Struct. Mol. Biol.* 2008. Vol. 15, № 1.
82. Meinhart A., Cramer P. Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors // *Nature*. 2004. Vol. 430, № 6996.
83. Xiang K. et al. Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex // *Nature*. 2010. Vol. 467, № 7316.
84. de Almeida S.F., Carmo-Fonseca M. Reciprocal regulatory links between cotranscriptional splicing and chromatin // *Seminars in Cell and Developmental Biology*. 2014. Vol. 32.
85. Parua P.K. et al. A Cdk9-PP1 switch regulates the elongation-termination transition of RNA polymerase II // *Nature*. 2018. Vol. 558, № 7710.
86. Nedeia E. et al. The Glc7 Phosphatase Subunit of the Cleavage and Polyadenylation Factor Is Essential for Transcription Termination on snoRNA Genes // *Mol. Cell*. 2008. Vol. 29, № 5.
87. Schrieck A. et al. RNA polymerase II termination involves C-terminal-domain tyrosine dephosphorylation by CPF subunit Glc7 // *Nat. Struct. Mol. Biol.* 2014. Vol. 21, № 2.
88. Zhang Y. et al. Structural Insights into the Human Pre-mRNA 3'-End Processing Machinery // *Mol. Cell*. 2020. Vol. 77, № 4.
89. Sun Y. et al. Molecular basis for the recognition of the human AAUAAA polyadenylation signal // *Proc. Natl. Acad. Sci. U. S. A.* 2018. Vol. 115, № 7.
90. Clerici M. et al. Structural insights into the assembly and polyA signal recognition mechanism of the human CPSF complex // *Elife*. 2017. Vol. 6.
91. Hamilton K., Sun Y., Tong L. Biophysical characterizations of the recognition of the AAUAAA polyadenylation signal // *RNA*. 2019. Vol. 25, № 12.
92. Kaufmann I. et al. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase // *EMBO J.* 2004. Vol. 23, № 3.

93. Topalian S.L. et al. Identification and Functional Characterization of Neo-Poly(A) Polymerase, an RNA Processing Enzyme Overexpressed in Human Tumors // *Mol. Cell. Biol.* 2001. Vol. 21, № 16.
94. Yang Q. et al. Crystal structure of human poly(A) polymerase gamma reveals a conserved catalytic core for canonical poly(A) polymerases // *J. Mol. Biol.* 2014. Vol. 426, № 1.
95. Lee Y.J., Lee Y., Chung J.H. An intronless gene encoding a poly(A) polymerase is specifically expressed in testis // *FEBS Lett.* 2000. Vol. 487, № 2.
96. Wahle E. Purification and characterization of a mammalian polyadenylate polymerase involved in the 3' end processing of messenger RNA precursors // *J. Biol. Chem.* 1991. Vol. 266, № 5.
97. Murthy K.G.K., Manley J.L. The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation // *Genes Dev.* 1995. Vol. 9, № 21.
98. Dominski Z. et al. A CPSF-73 Homologue Is Required for Cell Cycle Progression but Not Cell Growth and Interacts with a Protein Having Features of CPSF-100 // *Mol. Cell. Biol.* 2005. Vol. 25, № 4.
99. Dominski Z. The hunt for the 3' endonuclease // *Wiley Interdiscip. Rev. RNA.* 2010. Vol. 1, № 2.
100. Callebaut I. et al. Metallo- β -lactamase fold within nucleic acids processing enzymes: The β -CASP family // *Nucleic Acids Res.* 2002. Vol. 30, № 16.
101. Wu X., Bartel D.P. Widespread Influence of 3'-End Structures on Mammalian mRNA Processing and Stability // *Cell.* 2017. Vol. 169, № 5. P. 905-917.e11.
102. Sheets M.D., Ogg S.C., Wickens M.P. Point mutations in AAUAAA and the poly (A) addition site: Effects on the accuracy and efficiency of cleavage and polyadenylation in vitro // *Nucleic Acids Res.* 1990. Vol. 18, № 19.
103. Chen F., Macdonald C.C., Wilusf J. Cleavage site determinants in the mammalian polydenylation signal // *Nucleic Acids Res.* 1995. Vol. 23, № 14. P. 2614–2620.

104. Wang R. et al. A compendium of conserved cleavage and polyadenylation events in mammalian genes // *Genome Res.* 2018. Vol. 28, № 10.
105. Sun Y. et al. Structure of an active human histone pre-mRNA 3'-end processing machinery // *Science* (80-.). 2020. Vol. 367, № 6478.
106. Gehring N.H. et al. Increased efficiency of mRNA 3' end formation: A new genetic mechanism contributing to hereditary thrombophilia // *Nat. Genet.* 2001. Vol. 28, № 4.
107. Danckwardt S., Hentze M.W., Kulozik A.E. 3' end mRNA processing: Molecular mechanisms and implications for health and disease // *EMBO Journal.* 2008. Vol. 27, № 3.
108. Mandel C.R. et al. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease // *Nature.* 2006. Vol. 444, № 7121.
109. Liu H., Moore C.L. On the Cutting Edge: Regulation and Therapeutic Potential of the mRNA 3' End Nuclease // *Trends in Biochemical Sciences.* 2021. Vol. 46, № 9.
110. Liu H., Heller-Trulli D., Moore C.L. Targeting the mRNA endonuclease CPSF73 inhibits breast cancer cell migration, invasion, and self-renewal // *iScience.* 2022. Vol. 25, № 8.
111. Hunt A.G. et al. Transcriptional dynamics in the protozoan parasite *Sarcocystis* neurona and mammalian host cells after treatment with a specific inhibitor of apicomplexan mRNA polyadenylation // *PLoS One.* 2021. Vol. 16, № 10 October.
112. Kakegawa J. et al. JTE-607, a multiple cytokine production inhibitor, targets CPSF3 and inhibits pre-mRNA processing // *Biochem. Biophys. Res. Commun.* 2019. Vol. 518, № 1.
113. Ross N.T. et al. CPSF3-dependent pre-mRNA processing as a druggable node in AML and Ewing's sarcoma // *Nat. Chem. Biol.* 2020. Vol. 16, № 1.
114. Palencia A. et al. Targeting *Toxoplasma gondii* CPSF 3 as a new approach to control toxoplasmosis // *EMBO Mol. Med.* 2017. Vol. 9, № 3.
115. Sonoiki E. et al. A potent antimalarial benzoxaborole targets a Plasmodium

- falciparum cleavage and polyadenylation specificity factor homologue // Nat. Commun. 2017. Vol. 8.
116. Waithaka A., Clayton C. Clinically relevant benzoxaboroles inhibit mRNA processing in *Trypanosoma brucei* // BMC Res. Notes. 2022. Vol. 15, № 1.
 117. Jenny A. et al. Sequence similarity between the 73-kilodalton protein of mammalian CPSF and a subunit of yeast polyadenylation factor I // Science (80-.). 1996. Vol. 274, № 5292.
 118. Kennedy S.A. et al. Crystal Structure of the HEAT Domain from the Pre-mRNA Processing Factor Symplekin // J. Mol. Biol. 2009. Vol. 392, № 1.
 119. Ghazy M.A. et al. The Essential N Terminus of the Pta1 Scaffold Protein Is Required for snoRNA Transcription Termination and Ssu72 Function but Is Dispensable for Pre-mRNA 3'-End Processing // Mol. Cell. Biol. 2009. Vol. 29, № 8.
 120. Xiang K., Manley J.L., Tong L. An unexpected binding mode for a Pol II CTD peptide phosphorylated at Ser7 in the active site of the CTD phosphatase Ssu72 // Genes Dev. 2012. Vol. 26, № 20.
 121. Ruepp M.D. et al. Interactions of CstF-64, CstF-77, and symplekin: Implications on localization and function // Mol. Biol. Cell. 2011. Vol. 22, № 1.
 122. Krishnamurthy S. et al. Ssu72 is an RNA polymerase II CTD phosphatase // Mol. Cell. 2004. Vol. 14, № 3.
 123. Hausmann S. et al. Different strategies for carboxyl-terminal domain (CTD) recognition by serine 5-specific CTD phosphatases // J. Biol. Chem. 2005. Vol. 280, № 45.
 124. Werner-Allen J.W. et al. cis-proline-mediated ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72 // J. Biol. Chem. 2011. Vol. 286, № 7.
 125. Zhang D.W. et al. Ssu72 phosphatase-dependent erasure of phospho-Ser7 marks on the RNA polymerase II C-terminal domain is essential for viability and transcription termination // J. Biol. Chem. 2012. Vol. 287, № 11.

126. Deka P. et al. Protein and RNA dynamics play key roles in determining the specific recognition of GU-rich polyadenylation regulatory elements by human Cstf-64 protein // J. Mol. Biol. 2005. Vol. 347, № 4. P. 719–733.
127. Pérez Cañadillas J.M., Varani G. Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein // EMBO J. 2003. Vol. 22, № 11. P. 2821–2830.
128. Takagaki Y., Manley J.L. RNA Recognition by the Human Polyadenylation Factor CstF // Mol. Cell. Biol. 1997. Vol. 17, № 7. P. 3907–3914.
129. Beyer K., Dandekar T., Keller W. RNA ligands selected by cleavage stimulation factor contain distinct sequence motifs that function as downstream elements in 3'-end processing of pre-mRNA // J. Biol. Chem. © 1997 ASBMB. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology., 1997. Vol. 272, № 42. P. 26769–26779.
130. Monarez R.R., MacDonald C.C., Dass B. Polyadenylation proteins CstF-64 and τ CstF-64 exhibit differential binding affinities for RNA polymers // Biochem. J. 2007. Vol. 401, № 3. P. 651–658.
131. Wallace A.M. et al. Two distinct forms of the 64,000 Mr protein of the cleavage stimulation factor are expressed in mouse male germ cells // Proc. Natl. Acad. Sci. U. S. A. 1999. Vol. 96, № 12.
132. Li W. et al. The τ CstF-64 Polyadenylation Protein Controls Genome Expression in Testis // PLoS One. 2012. Vol. 7, № 10.
133. Yao C. et al. Overlapping and distinct functions of CstF64 and CstF64 τ in mammalian mRNA 3' processing // RNA. 2013. Vol. 19, № 12.
134. Takagaki Y., Manley J.L. Complex Protein Interactions within the Human Polyadenylation Machinery Identify a Novel Component // Mol. Cell. Biol. 2000. Vol. 20, № 5.
135. Gross S., Moore C. Five subunits are required for reconstitution of the cleavage and polyadenylation activities of *Saccharomyces cerevisiae* cleavage factor I // Proc. Natl. Acad. Sci. U. S. A. 2001. Vol. 98, № 11.

136. Chennathukuzhi V.M. et al. Elevated levels of the polyadenylation factor CstF 64 enhance formation of the 1kB Testis Brain RNA-Binding Protein (TB-RBP) mRNA in male germ cells // *Mol. Reprod. Dev.* 2001. Vol. 58, № 4.
137. Colgan D.F., Manley J.L. Mechanism and regulation of mRNA polyadenylation // *Genes Dev.* 1997. Vol. 11, № 21. P. 2755–2766.
138. Bai Y. et al. Crystal Structure of Murine CstF-77: Dimeric Association and Implications for Polyadenylation of mRNA Precursors // *Mol. Cell.* 2007. Vol. 25, № 6.
139. Hockert J.A., Yeh H.J., MacDonald C.C. The Hinge domain of the cleavage stimulation factor protein CstF-64 is essential for CstF-77 interaction, nuclear localization, and polyadenylation // *J. Biol. Chem.* 2010. Vol. 285, № 1.
140. Takagaki Y., Manley J.L. A polyadenylation factor subunit is the human homologue of the *Drosophila* suppressor of forked protein // *Nature.* 1994. Vol. 372, № 6505.
141. Legrand P. et al. The structure of the CstF-77 homodimer provides insights into CstF assembly // *Nucleic Acids Res.* 2007. Vol. 35, № 13.
142. Moreno-Morcillo M. et al. Hexameric architecture of CstF supported by CstF-50 homodimerization domain structure // *RNA.* 2011. Vol. 17, № 3.
143. Kleiman F.E., Manley J.L. Functional interaction of BRCA1-associated BARD1 with polyadenylation factor CstF-50 // *Science (80-.).* 1999. Vol. 285, № 5433.
144. Kleiman F.E., Manley J.L. The BARD1-CstF-50 interaction links mRNA 3' end formation to DNA damage and tumor suppression // *Cell.* 2001. Vol. 104, № 5.
145. Schäfer P. et al. Reconstitution of mammalian cleavage factor II involved in 3' processing of mRNA precursors // *RNA.* 2018. Vol. 24, № 12.
146. Zhang Z., Fu J., Gilmour D.S. CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3'-end processing factor, Pcf11 // *Genes Dev.* 2005. Vol. 19, № 13.
147. Ghazy M.A. et al. The interaction of Pcf11 and Clp1 is needed for mRNA 3'-end

- formation and is modulated by amino acids in the ATP-binding site // *Nucleic Acids Res.* 2012. Vol. 40, № 3.
148. Sadowski M. et al. Independent functions of yeast Pcf11p in pre-mRNA 3' end processing and in transcription termination // *EMBO J.* 2003. Vol. 22, № 9.
 149. Barillà D., Lee B.A., Proudfoot N.J. Cleavage/polyadenylation factor IA associates with the carboxyl-terminal domain of RNA polymerase II in *Saccharomyces cerevisiae* // *Proc. Natl. Acad. Sci. U. S. A.* 2001. Vol. 98, № 2.
 150. Licatalosi D.D. et al. Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II // *Mol. Cell.* 2002. Vol. 9, № 5.
 151. Komarnitsky P., Cho E.J., Buratowski S. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription // *Genes Dev.* 2000. Vol. 14, № 19.
 152. Weitzer S., Martinez J. The human RNA kinase hClp1 is active on 3' transfer RNA exons and short interfering RNAs // *Nature.* 2007. Vol. 447, № 7141.
 153. De Vries H. et al. Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors // *EMBO J.* 2000. Vol. 19, № 21.
 154. Noble C.G., Beuth B., Taylor I.A. Structure of a nucleotide-bound Clp1-Pcf11 polyadenylation factor // *Nucleic Acids Res.* 2007. Vol. 35, № 1.
 155. Walker J.E. et al. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. // *EMBO J.* 1982. Vol. 1, № 8.
 156. Ramirez A., Shuman S., Schwer B. Human RNA 5'-kinase (hClp1) can function as a tRNA splicing enzyme in vivo // *RNA.* 2008. Vol. 14, № 9.
 157. Schmidt M. et al. Reconstitution of 3' end processing of mammalian pre-mRNA reveals a central role of RBBP6 // *Genes Dev.* 2022. Vol. 36, № 3–4. P. 195–209.
 158. Haddad R. et al. An essential role for Clp1 in assembly of polyadenylation complex CF IA and Pol II transcription termination // *Nucleic Acids Res.* 2012.

159. Holbein S. et al. The P-loop domain of yeast Clp1 mediates interactions between CF IA and CPF factors in pre-mRNA 3' end formation // PLoS One. 2011. Vol. 6, № 12.
160. Boreikaite V. et al. RBBP6 activates the pre-mRNA 3' end processing machinery in humans // Genes Dev. 2022. Vol. 36, № 3–4.
161. Hill C.H. et al. Activation of the Endonuclease that Defines mRNA 3' Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex // Mol. Cell. 2019. Vol. 73, № 6.
162. Rodríguez-Molina J.B. et al. Mpe1 senses the binding of pre-mRNA and controls 3' end processing by CPF // Mol. Cell. 2022. Vol. 82, № 13.
163. Di Giammartino D.C. et al. RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3' UTRs // Genes Dev. 2014. Vol. 28, № 20. P. 2248–2260.
164. Gruber A.J., Zavolan M. Alternative cleavage and polyadenylation in health and disease // Nature Reviews Genetics. 2019. Vol. 20, № 10.
165. Hu J. et al. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation // Rna. 2005. Vol. 11, № 10. P. 1485–1493.
166. Hwang H.W. et al. PAPERCLIP Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage // Cell Reports. 2016. Vol. 15, № 2. P. 423–435.
167. Masamha C.P. et al. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression // Nature. 2014. Vol. 510, № 7505.
168. Gruber A.R. et al. Cleavage factor Im is a key regulator of 3' UTR length // RNA Biol. 2012. Vol. 9, № 12.
169. Kubo T. et al. Knock-down of 25 kDa subunit of cleavage factor Im in HeLa cells alters alternative polyadenylation within 3'-UTRs // Nucleic Acids Res. 2006. Vol.

34, № 21.

170. Neve J. et al. Cleavage and polyadenylation: Ending the message expands gene regulation // *RNA Biology*. 2017. Vol. 14, № 7. P. 865–890.
171. Li W. et al. Systematic Profiling of Poly(A)+ Transcripts Modulated by Core 3' End Processing and Splicing Factors Reveals Regulatory Rules of Alternative Cleavage and Polyadenylation // *PLoS Genet*. 2015. Vol. 11, № 4.
172. Stroup E.K., Ji Z. Deep learning of human polyadenylation sites at nucleotide resolution reveals molecular determinants of site usage and relevance in disease // *Nat. Commun*. 2023. Vol. 14, № 1.
173. Oikonomou P., Goodarzi H., Tavazoie S. Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts // *Cell Reports*. 2014. Vol. 7, № 1. P. 281–292.
174. Gamberi C. et al. An anterior function for the *Drosophila* posterior determinant Pumilio // *Development*. 2002. Vol. 129, № 11.
175. Goldstrohm A.C., Hall T.M.T., McKenney K.M. Post-transcriptional Regulatory Functions of Mammalian Pumilio Proteins // *Trends in Genetics*. 2018. Vol. 34, № 12.
176. Wang X., Zamore P.D., Tanaka Hall T.M. Crystal structure of a Pumilio homology domain // *Mol. Cell*. 2001. Vol. 7, № 4.
177. Wickens M. et al. A PUF family portrait: 3'UTR regulation as a way of life // *Trends in Genetics*. 2002. Vol. 18, № 3.
178. Goldstrohm A.C. et al. PUF proteins bind Pop2p to regulate messenger RNAs // *Nat. Struct. Mol. Biol*. 2006. Vol. 13, № 6.
179. Martin G. et al. Genome-wide Analysis of Pre-mRNA 3' End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3' UTR Length // *Cell Rep*. 2012. Vol. 1, № 6.
180. Kim S. et al. Evidence that cleavage factor Im is a heterotetrameric protein complex controlling alternative polyadenylation // *Genes to Cells*. 2010. Vol. 15,

181. Brown K.M., Gilmartin G.M. A Mechanism for the Regulation of Pre-mRNA 3' Processing by Human Cleavage Factor Im // *Mol. Cell*. 2003. Vol. 12, № 6.
182. Ruepp M.D., Schümperli D., Barabino S.M.L. mRNA 3' end processing and more-multiple functions of mammalian cleavage factor I-68 // *Wiley Interdisciplinary Reviews: RNA*. 2011. Vol. 2, № 1.
183. Zhu Y. et al. Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation // *Mol. Cell*. 2018. Vol. 69, № 1.
184. Chao L.C. et al. Assembly of the Cleavage and Polyadenylation Apparatus Requires About 10 Seconds In Vivo and Is Faster for Strong than for Weak Poly(A) Sites // *Mol. Cell. Biol*. 1999. Vol. 19, № 8.
185. Klasens B.I.F., Das A.T., Berkhout B. Inhibition of polyadenylation by stable RNA secondary structure // *Nucleic Acids Res*. 1998. Vol. 26, № 8. P. 1870–1876.
186. Wu C., Alwine J.C. Secondary Structure as a Functional Feature in the Downstream Region of Mammalian Polyadenylation Signals // *Mol. Cell. Biol*. 2004. Vol. 24, № 7. P. 2789–2796.
187. Debashish Ray et al. A compendium of RNA-binding motifs for decoding gene regulation // *Nature*. 2013. Vol. 499, № 7457. P. 172–177.
188. Dubbury S.J., Boutz P.L., Sharp P.A. CDK12 regulates DNA repair genes by suppressing intronic polyadenylation // *Nature*. 2018. Vol. 564, № 7734.
189. Gregersen L.H. et al. SCAF4 and SCAF8, mRNA Anti-Terminator Proteins // *Cell*. 2019. Vol. 177, № 7.
190. Kaida D. et al. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation // *Nature*. 2010. Vol. 468, № 7324.
191. Oh J.M. et al. U1 snRNP telescripting regulates a size-function-stratified human genome // *Nat. Struct. Mol. Biol*. 2017. Vol. 24, № 11.
192. Berg M.G. et al. U1 snRNP determines mRNA length and regulates isoform expression // *Cell*. 2012. Vol. 150, № 1.

193. Almada A.E. et al. Promoter directionality is controlled by U1 snRNP and polyadenylation signals // *Nature*. 2013. Vol. 499, № 7458.
194. Niwa M., Berget S.M. Polyadenylation precedes splicing in vitro. // *Gene Expr.* 1991. Vol. 1, № 1.
195. Niwa M., MacDonald C.C., Berget S.M. Are vertebrate exons scanned during splice-site selection? // *Nature*. 1992. Vol. 360, № 6401.
196. Dye M.J., Proudfoot N.J. Terminal exon definition occurs cotranscriptionally and promotes termination of RNA polymerase II // *Mol. Cell*. 1999. Vol. 3, № 3.
197. Vagner S., Vagner C., Mattaj I.W. The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3'-end processing and splicing // *Genes Dev.* 2000. Vol. 14, № 4.
198. Kyburz A. et al. Direct Interactions between Subunits of CPSF and the U2 snRNP Contribute to the Coupling of Pre-mRNA 3' End Processing and Splicing // *Mol. Cell*. 2006. Vol. 23, № 2.
199. Bienroth S., Keller W., Wahle E. Assembly of a processive messenger RNA polyadenylation complex. // *EMBO J.* 1993. Vol. 12, № 2.
200. Nemeth A. et al. Isolation of genomic and cDNA clones encoding bovine poly(A) binding protein II // *Nucleic Acids Res.* 1995. Vol. 23, № 20.
201. Wahle E. Poly(A) tail length control is caused by termination of processive synthesis // *J. Biol. Chem.* 1995. Vol. 270, № 6.
202. Kerwitz Y. et al. Stimulation of poly(A) polymerase through a direct interaction with the nuclear poly(A) binding protein allosterically regulated by RNA // *EMBO J.* 2003. Vol. 22, № 14.
203. Wahle E. A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation // *Cell*. 1991. Vol. 66, № 4.
204. Kühn U. et al. Poly(A) tail length is controlled by the nuclear Poly(A)-binding protein regulating the interaction between Poly(A) polymerase and the cleavage and polyadenylation specificity factor // *J. Biol. Chem.* 2009. Vol. 284, № 34.

205. Edmonds M., Vaughan M.H., Nakazato H. Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship. // Proc. Natl. Acad. Sci. U. S. A. 1971. Vol. 68, № 6.
206. McLaughlin C.S. et al. Polyadenylic acid sequences in yeast messenger ribonucleic acid. // J. Biol. Chem. 1973. Vol. 248, № 4.
207. Brown C.E., Sachs A.B. Poly(A) Tail Length Control in *Saccharomyces cerevisiae* Occurs by Message-Specific Deadenylation // Mol. Cell. Biol. 1998. Vol. 18, № 11.
208. Chen Z., Li Y., Krug R.M. Influenza A virus NS1 protein targets poly(A)-binding protein II of the cellular 3'-end processing machinery // EMBO J. 1999. Vol. 18, № 8.
209. Apponi L.H. et al. Loss of nuclear poly(A)-binding protein 1 causes defects in myogenesis and mRNA biogenesis // Hum. Mol. Genet. 2010. Vol. 19, № 6.
210. Saguez C. et al. Nuclear mRNA Surveillance in THO/sub2 Mutants Is Triggered by Inefficient Polyadenylation // Mol. Cell. 2008. Vol. 31, № 1.
211. Jensen T.H., Jacquier A., Libri D. Dealing with pervasive transcription // Molecular Cell. 2013. Vol. 52, № 4.
212. Gudipati R.K. et al. Extensive Degradation of RNA Precursors by the Exosome in Wild-Type Cells // Mol. Cell. 2012. Vol. 48, № 3.
213. Tang P. et al. Alternative polyadenylation by sequential activation of distal and proximal PolyA sites // Nature Structural and Molecular Biology. 2022. Vol. 29, № 1. P. 21–31.
214. Hao Y. et al. Sequential Polyadenylation to Enable Alternative mRNA 3' End Formation // Mol. Cells. 2023. Vol. 46, № 1. P. 57–64.
215. Tang P. et al. Alternative polyadenylation by sequential activation of distal and proximal PolyA sites // Nat. Struct. Mol. Biol. Springer US, 2022. Vol. 29, № 1. P. 21–31.

216. Gallie D.R. The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency // *Genes Dev.* 1991. Vol. 5, № 11.
217. Baer B.W., Kornberg R.D. The protein responsible for the repeating structure of cytoplasmic poly(A)-ribonucleoprotein // *J. Cell Biol.* 1983. Vol. 96, № 3.
218. Chang H. et al. TAIL-seq: Genome-wide determination of poly(A) tail length and 3' end modifications // *Mol. Cell.* 2014. Vol. 53, № 6.
219. Eisen T.J. et al. The Dynamics of Cytoplasmic mRNA Metabolism // *Mol. Cell.* 2020. Vol. 77, № 4.
220. Cheng J. et al. Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast // *Rna.* 2017. Vol. 23, № 11. P. 1648–1659.
221. Tanguay R.L., Gallie D.R. Translational Efficiency Is Regulated by the Length of the 3' Untranslated Region // *Mol. Cell. Biol.* 1996. Vol. 16, № 1.
222. Jackson R.J., Hellen C.U.T., Pestova T. V. The mechanism of eukaryotic translation initiation and principles of its regulation // *Nature Reviews Molecular Cell Biology.* 2010. Vol. 11, № 2.
223. Chen J., Kastan M.B. 5'-3'-UTR interactions regulate p53 mRNA translation and provide a target for modulating p53 induction after DNA damage // *Genes Dev.* 2010. Vol. 24, № 19.
224. Mazumder B. et al. Regulated release of L13a from the 60S ribosomal subunit as a mechanism of transcript-specific translational control // *Cell.* 2003. Vol. 115, № 2.
225. Passmore L.A., Collier J. Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression // *Nature Reviews Molecular Cell Biology.* 2022. Vol. 23, № 2. P. 93–106.
226. Charlesworth A. et al. Musashi regulates the temporal order of mRNA translation during *Xenopus* oocyte maturation // *EMBO J.* 2006. Vol. 25, № 12.
227. Radford H.E., Meijer H.A., de Moor C.H. Translational control by cytoplasmic polyadenylation in *Xenopus* oocytes // *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms.* 2008. Vol. 1779, № 4.

228. Charlesworth A., Meijer H.A., De Moor C.H. Specificity factors in cytoplasmic polyadenylation // Wiley Interdiscip. Rev. RNA. 2013. Vol. 4, № 4.
229. Mellman D.L. et al. A PtdIns4,5P2-regulated nuclear poly(A) polymerase controls expression of select mRNAs // Nature. 2008. Vol. 451, № 7181.
230. Yamashita S. et al. Crystal structures of U6 snRNA-specific terminal uridylyltransferase // Nat. Commun. 2017. Vol. 8.
231. Trippe R. et al. Identification, cloning, and functional analysis of the human U6 snRNA-specific terminal uridylyl transferase // RNA. 2006. Vol. 12, № 8.
232. Kandala D.T. et al. CstF-64 and 3'-UTR cis-element determine Star-PAP specificity for target mRNA selection by excluding PAP α // Nucleic Acids Res. 2015. Vol. 44, № 2. P. 811–823.
233. Laishram R.S., Anderson R.A. The poly A polymerase Star-PAP controls 3'-2'-end cleavage by promoting CPSF interaction and specificity toward the pre-mRNA // EMBO J. 2010. Vol. 29, № 24.
234. Li W. et al. Star-PAP control of BIK expression and apoptosis is regulated by nuclear PIPKI α and PKC δ signaling // Mol. Cell. 2012. Vol. 45, № 1.
235. Li W. et al. Distinct regulation of alternative polyadenylation and gene expression by nuclear poly(A) polymerases // Nucleic Acids Res. Oxford University Press, 2017. Vol. 45, № 15. P. 8930–8942.
236. Sudheesh A.P. et al. Star-PAP controlled alternative polyadenylation coupled poly(A) tail length regulates protein expression in hypertrophic heart // Nucleic Acids Res. 2019. Vol. 47, № 20.
237. Duan A. et al. Star-PAP regulates tumor protein D52 through modulating miR-449a/34a in breast cancer // Biol. Open. 2019. Vol. 8, № 11.
238. Yu C. et al. Star-PAP, a poly(A) polymerase, functions as a tumor suppressor in an orthotopic human breast cancer model // Cell Death Dis. 2017. Vol. 8, № 2.
239. A.P. S., Laishram R.S. Nuclear Phosphatidylinositol-Phosphate Type I Kinase α -Coupled Star-PAP Polyadenylation Regulates Cell Invasion // Mol. Cell. Biol.

2018. Vol. 38, № 5.

240. Laishram R.S., Barlow C.A., Anderson R.A. CKI isoforms α and ϵ regulate Star-PAP target messages by controlling Star-PAP poly(A) polymerase activity and phosphoinositide stimulation // *Nucleic Acids Res.* 2011. Vol. 39, № 18.
241. Gonzales M.L., Mellman D.L., Anderson R.A. CKI α is associated with and phosphorylates Star-PAP and is also required for expression of select Star-PAP target messenger RNAs // *J. Biol. Chem.* 2008. Vol. 283, № 18.
242. Derti A. et al. A quantitative atlas of polyadenylation in five mammals // *Genome Res.* 2012. Vol. 22, № 6.
243. Tian B. et al. A large-scale analysis of mRNA polyadenylation of human and mouse genes // *Nucleic Acids Res.* 2005. Vol. 33, № 1. P. 201–212.
244. Bartel D.P. MicroRNAs: Target Recognition and Regulatory Functions // *Cell.* 2009. Vol. 136, № 2.
245. Chen C.Y. et al. AU binding proteins recruit the exosome to degrade ARE-containing mRNAs // *Cell.* 2001. Vol. 107, № 4.
246. Tran H. et al. Facilitation of mRNA Deadenylation and Decay by the Exosome-Bound, DExH Protein RHAU // *Mol. Cell.* 2004. Vol. 13, № 1.
247. Lykke-Andersen J., Wagner E. Recruitment and activation of mRNA decay enzymes by two ARE-mediated decay activation domains in the proteins TTP and BRF-1 // *Genes Dev.* 2005. Vol. 19, № 3.
248. Chamboredon S. et al. Hypoxia-inducible factor-1 α mRNA: A new target for destabilization by tristetraprolin in endothelial cells // *Mol. Biol. Cell.* 2011. Vol. 22, № 18.
249. Knapinska A.M. et al. Chaperone Hsp27 Modulates AUF1 Proteolysis and AU-Rich Element-Mediated mRNA Degradation // *Mol. Cell. Biol.* 2011. Vol. 31, № 7.
250. Berkovits B.D., Mayr C. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization // *Nature.* 2015. Vol. 522, № 7556.
251. St. Johnston D. Moving messages: The intracellular localization of mRNAs //

- Nature Reviews Molecular Cell Biology. 2005. Vol. 6, № 5.
252. Lutz C.S., Moreira A. Alternative mRNA polyadenylation in eukaryotes: An effective regulator of gene expression // Wiley Interdisciplinary Reviews: RNA. 2011. Vol. 2, № 1.
 253. Taliaferro J.M. et al. Distal Alternative Last Exons Localize mRNAs to Neural Projections // Mol. Cell. 2016. Vol. 61, № 6.
 254. Geisberg J. V. et al. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast // Cell. Elsevier Inc., 2014. Vol. 156, № 4. P. 812–824.
 255. De Klerk E. et al. Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation // Nucleic Acids Res. 2012. Vol. 40, № 18.
 256. Morris A.R. et al. Alternative cleavage and polyadenylation during colorectal cancer development // Clin. Cancer Res. 2012. Vol. 18, № 19.
 257. Elkon R. et al. E2F mediates enhanced alternative polyadenylation in proliferation // Genome Biol. 2012. Vol. 13, № 7.
 258. Neve J., Furger A. Alternative polyadenylation: Less than meets the eye? // Biochemical Society Transactions. 2014. Vol. 42, № 4.
 259. Fu Y. et al. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing // Genome Res. 2011. Vol. 21, № 5.
 260. Neve J. et al. Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation // Genome Res. 2016. Vol. 26, № 1.
 261. Gruber A.R. et al. Means to an end: Mechanisms of alternative polyadenylation of messenger RNA precursors // Wiley Interdisciplinary Reviews: RNA. 2014. Vol. 5, № 2.
 262. Gupta I. et al. Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions // Mol. Syst. Biol. 2014. Vol. 10, № 2.

263. Alt F.W. et al. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends // *Cell*. 1980. Vol. 20, № 2.
264. Takagaki Y. et al. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation // *Cell*. 1996. Vol. 87, № 5.
265. Takagaki Y., Manley J.L. Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation // *Mol. Cell*. 1998. Vol. 2, № 6.
266. Ashfield R., Enriquez-Harris P., Proudfoot N.J. Transcriptional termination between the closely linked human complement genes C2 and factor B: common termination factor for C2 and c-myc? // *EMBO J*. 1991. Vol. 10, № 13.
267. Fong N. et al. Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition // *Mol. Cell*. 2015. Vol. 60, № 2.
268. Jenal M. et al. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites // *Cell*. 2012. Vol. 149, № 3.
269. Flavell S.W. et al. Genome-Wide Analysis of MEF2 Transcriptional Program Reveals Synaptic Target Genes and Neuronal Activity-Dependent Polyadenylation Site Selection // *Neuron*. 2008. Vol. 60, № 6.
270. Ji Z. et al. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development // *Proc. Natl. Acad. Sci. U. S. A.* 2009. Vol. 106, № 17.
271. Ji Z., Tian B. Reprogramming of 3' Untranslated Regions of mRNAs by Alternative Polyadenylation in Generation of Pluripotent Stem Cells from Different Cell Types // *PLoS One*. 2009. Vol. 4, № 12.
272. Li W. et al. Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control // *BMC Biol. BMC Biology*, 2016. Vol. 14, № 1. P. 1–17.

273. Liu D. et al. Systematic variation in mRNA 3'-processing signals during mouse spermatogenesis // *Nucleic Acids Res.* 2007. Vol. 35, № 1.
274. Miura P. et al. Widespread and extensive lengthening of 3' UTRs in the mammalian brain // *Genome Res.* 2013. Vol. 23, № 5.
275. Moqtaderi Z. et al. Species-specific factors mediate extensive heterogeneity of mRNA 3' ends in yeasts // *Proc. Natl. Acad. Sci. U. S. A.* 2013. Vol. 110, № 27.
276. Velten L. et al. Single- cell polyadenylation site mapping reveals 3' isoform choice variability // *Mol. Syst. Biol.* 2015. Vol. 11, № 6.
277. Logan J. et al. A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. // *Proc. Natl. Acad. Sci. U. S. A.* 1987. Vol. 84, № 23.
278. Connelly S., Manley J.L. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. // *Genes Dev.* 1988. Vol. 2, № 4. P. 440–452.
279. Proudfoot N.J. How RNA polymerase II terminates transcription in higher eukaryotes // *Trends in Biochemical Sciences.* 1989. Vol. 14, № 3.
280. Eaton J.D., West S. Termination of Transcription by RNA Polymerase II: BOOM! // *Trends Genet.* Elsevier Ltd, 2020. Vol. 36, № 9. P. 664–675.
281. Kim M. et al. The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II // *Nature.* 2004. Vol. 432, № 7016.
282. West S., Gromak N., Proudfoot N.J. Human 5' → 3' exonuclease Xpm1 promotes transcription termination at co-transcriptional cleavage sites // *Nature.* 2004. Vol. 432, № 7016. P. 522–525.
283. Luo W., Johnson A.W., Bentley D.L. The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: Implications for a unified allosteric-torpedo model // *Genes Dev.* 2006. Vol. 20, № 8.
284. Cortazar M.A. et al. Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a “Sitting Duck Torpedo”

Mechanism // Mol. Cell. 2019. Vol. 76, № 6.

- 285. Eaton J.D. et al. A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes // Genes Dev. 2020. Vol. 34, № 1–2.
- 286. Nojima T. et al. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing // Cell. 2015. Vol. 161, № 3.
- 287. Kecman T. et al. Elongation/Termination Factor Exchange Mediated by PP1 Phosphatase Orchestrates Transcription Termination // Cell Rep. 2018. Vol. 25, № 1.
- 288. Gromak N., West S., Proudfoot N.J. Pause Sites Promote Transcriptional Termination of Mammalian RNA Polymerase II // Mol. Cell. Biol. 2006. Vol. 26, № 10. P. 3986–3996.
- 289. West S., Proudfoot N.J. Human Pcf11 enhances degradation of RNA polymerase II-associated nascent RNA and transcriptional termination // Nucleic Acids Res. 2008. Vol. 36, № 3.
- 290. Creamer T.J. et al. Transcriptome-wide binding sites for components of the *Saccharomyces cerevisiae* non-poly(A) termination pathway: Nrd1, Nab3, and Sen1 // PLoS Genet. 2011. Vol. 7, № 10.
- 291. Baejen C. et al. Genome-wide Analysis of RNA Polymerase II Termination at Protein-Coding Genes // Mol. Cell. 2017. Vol. 66, № 1. P. 38-49.e6.
- 292. Birse C.E. et al. Coupling termination of transcription to messenger RNA maturation in yeast // Science (80-.). 1998. Vol. 280, № 5361.
- 293. Schwalb B. et al. TT-seq maps the human transient transcriptome // Science (80-.). 2016. Vol. 352, № 6290. P. 1225–1228.
- 294. Park J., Kang M., Kim M. Unraveling the mechanistic features of RNA polymerase II termination by the 5'-3' exoribonuclease Rat1 // Nucleic Acids Res. 2015. Vol. 43, № 5.
- 295. Lunde B.M. et al. Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain // Nat. Struct. Mol. Biol. 2010. Vol. 17,

№ 10.

296. Skourti-Stathaki K., Proudfoot N.J., Gromak N. Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination // *Mol. Cell.* 2011. Vol. 42, № 6.
297. Zhang H., Rigo F., Martinson H.G. Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site // *Mol. Cell.* Elsevier Inc., 2015. Vol. 59, № 3. P. 437–448.
298. Lee J.S., Mendell J.T. Antisense-Mediated Transcript Knockdown Triggers Premature Transcription Termination // *Mol. Cell.* 2020. Vol. 77, № 5.
299. Lai F. et al. Directed RNase H Cleavage of Nascent Transcripts Causes Transcription Termination // *Mol. Cell.* Elsevier Inc., 2020. Vol. 77, № 5. P. 1032-1043.e4.
300. Greger I.H., Proudfoot N.J. Poly(A) signals control both transcriptional termination and initiation between the tandem GAL10 and GAL7 genes of *Saccharomyces cerevisiae* // *EMBO J.* 1998. Vol. 17, № 16.
301. Shearwin K.E., Callen B.P., Egan J.B. Transcriptional interference - A crash course // *Trends in Genetics.* 2005. Vol. 21, № 6.
302. Prescott E.M., Proudfoot N.J. Transcriptional collision between convergent genes in budding yeast // *Proc. Natl. Acad. Sci. U. S. A.* 2002. Vol. 99, № 13.
303. Hobson D.J. et al. RNA Polymerase II Collision Interrupts Convergent Transcription // *Mol. Cell.* 2012. Vol. 48, № 3.
304. Gaillard H., García-Muse T., Aguilera A. Replication stress and cancer // *Nature Reviews Cancer.* 2015. Vol. 15, № 5.
305. Steinmetz E.J. et al. cis - and trans -Acting Determinants of Transcription Termination by Yeast RNA Polymerase II // *Mol. Cell. Biol.* 2006. Vol. 26, № 7.
306. Kim M. et al. Distinct Pathways for snoRNA and mRNA Termination // *Mol. Cell.* 2006. Vol. 24, № 5.

307. Lenstra T.L. et al. The role of Ctk1 kinase in termination of small non-coding RNAs // PLoS One. 2013. Vol. 8, № 12.
308. Ghazal G. et al. Yeast RNase III Triggers Polyadenylation-Independent Transcription Termination // Mol. Cell. 2009. Vol. 36, № 1.
309. Rondón A.G. et al. Fail-Safe Transcriptional Termination for Protein-Coding Genes in *S. cerevisiae* // Mol. Cell. 2009. Vol. 36, № 1.
310. Colin J. et al. Roadblock termination by reb1p restricts cryptic and readthrough transcription // Mol. Cell. 2014. Vol. 56, № 5.
311. Gudipati R.K. et al. The yeast RPL9B gene is regulated by modulation between two modes of transcription termination // EMBO J. 2012. Vol. 31, № 10.
312. White E. et al. AT-rich sequence elements promote nascent transcript cleavage leading to RNA polymerase II termination // Nucleic Acids Res. 2013. Vol. 41, № 3. P. 1797–1806.
313. Dye M.J., Proudfoot N.J. Multiple transcript cleavage precedes polymerase release in termination by RNA polymerase II // Cell. 2001. Vol. 105, № 5. P. 669–681.
314. West S., Proudfoot N.J. Transcriptional Termination Enhances Protein Expression in Human Cells // Mol. Cell. 2009. Vol. 33, № 3. P. 354–364.
315. Nojima T. et al. Definition of RNA Polymerase II CoTC Terminator Elements in the Human Genome // Cell Rep. The Authors, 2013. Vol. 3, № 4. P. 1080–1092.
316. Marzluff W.F., Koreski K.P. Birth and Death of Histone mRNAs // Trends in Genetics. 2017. Vol. 33, № 10.
317. Romeo V., Schümperli D. Cycling in the nucleus: Regulation of RNA 3' processing and nuclear organization of replication-dependent histone genes // Current Opinion in Cell Biology. 2016. Vol. 40.
318. Furger A., Schaller A., Schümperli D. Functional importance of conserved nucleotides at the histone RNA 3' processing site // RNA. 1998. Vol. 4, № 3.
319. Dominski Z., Marzluff W.F. Formation of the 3' end of histone mRNA: Getting closer to the end // Gene. 2007. Vol. 396, № 2.

320. Skrajna A. et al. Protein composition of catalytically active U7-dependent processing complexes assembled on histone pre-mRNA containing biotin and a photo-cleavable linker // *Nucleic Acids Res.* 2018. Vol. 46, № 9.
321. Yang X.-C. et al. A Complex Containing the CPSF73 Endonuclease and Other Polyadenylation Factors Associates with U7 snRNP and Is Recruited to Histone Pre-mRNA for 3'-End Processing // *Mol. Cell. Biol.* 2013. Vol. 33, № 1.
322. Dominski Z., Yang X.C., Marzluff W.F. The polyadenylation factor CPSF-73 is involved in histone-pre-mRNA processing // *Cell.* 2005. Vol. 123, № 1.
323. Sanford J.R., Caceres J.F. Pre-mRNA splicing: Life at the centre of the central dogma // *J. Cell Sci.* 2004. Vol. 117, № 26.
324. Zheng H. et al. Identification of Integrator-PP2A complex (INTAC), an RNA polymerase II phosphatase // *Science (80-.).* 2020. Vol. 370, № 6520.
325. Baillat D. et al. Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II // *Cell.* 2005. Vol. 123, № 2.
326. Guirio J., Murphy S. Regulation of expression of human RNA polymerase II-Transcribed snRNA genes // *Open Biology.* 2017. Vol. 7, № 6.
327. Pfeleiderer M.M., Galej W.P. Structure of the catalytic core of the Integrator complex // *Mol. Cell.* 2021. Vol. 81, № 6.
328. Fianu I. et al. Structural basis of Integrator-mediated transcription regulation // *Science (80-.).* 2021. Vol. 374, № 6569.
329. Lardelli R.M., Lykke-Andersen J. Competition between maturation and degradation drives human snRNA 3' end quality control // *Genes Dev.* 2020. Vol. 34, № 13–14.
330. Davidson L. et al. DNA-directed termination of mammalian RNA polymerase II // *bioRxiv.* 2024. Vol. 0044, № 0. P. 2024.06.01.596947.
331. Hernandez N., Weiner A.M. Formation of the 3' end of U1 snRNA requires compatible snRNA promoter elements // *Cell.* 1986. Vol. 47, № 2.

332. O'Reilly D. et al. Human snRNA genes use polyadenylation factors to promote efficient transcription termination // *Nucleic Acids Res.* 2014. Vol. 42, № 1.
333. Eaton J.D. et al. Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity // *Genes Dev.* 2018. Vol. 32, № 2.
334. Suzuki H. et al. The 3' Pol II pausing at replication-dependent histone genes is regulated by Mediator through Cajal bodies' association with histone locus bodies // *Nat. Commun.* 2022. Vol. 13, № 1.
335. Vasiljeva L., Buratowski S. Nrd1 interacts with the nuclear exosome for 3' processing of RNA polymerase II transcripts // *Mol. Cell.* 2006. Vol. 21, № 2.
336. Carroll K.L. et al. Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements // *RNA.* 2007. Vol. 13, № 3.
337. Porrua O. et al. In vivo SELEX reveals novel sequence and structural determinants of Nrd1-Nab3-Sen1-dependent transcription termination // *EMBO J.* 2012. Vol. 31, № 19.
338. Wlotzka W. et al. The nuclear RNA polymerase II surveillance system targets polymerase III transcripts // *EMBO J.* 2011. Vol. 30, № 9.
339. Heo D.H. et al. The RNA polymerase 2 C-terminal domain-interacting domain of yeast Nrd1 contributes to the choice of termination pathway and couples to RNA processing by the nuclear exosome // *J. Biol. Chem.* 2013. Vol. 288, № 51.
340. Kubicek K. et al. Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1 // *Genes Dev.* 2012. Vol. 26, № 17.
341. Tudek A. et al. Molecular basis for coordinating transcription termination with noncoding RNA degradation // *Mol. Cell.* 2014. Vol. 55, № 3.
342. Vasiljeva L. et al. The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain // *Nat. Struct. Mol. Biol.* 2008. Vol. 15, № 8.
343. Porrua O., Libri D. A bacterial-like mechanism for transcription termination by the Sen1p helicase in budding yeast // *Nat. Struct. Mol. Biol.* 2013. Vol. 20, № 7.

344. Allmang C. et al. Functions of the exosome in rRNA, snoRNA and snRNA synthesis // EMBO J. 1999. Vol. 18, № 19.
345. van Hoof A., Lennertz P., Parker R. Yeast Exosome Mutants Accumulate 3'-Extended Polyadenylated Forms of U4 Small Nuclear RNA and Small Nucleolar RNAs // Mol. Cell. Biol. 2000. Vol. 20, № 2.
346. Arigo J.T. et al. Termination of Cryptic Unstable Transcripts Is Directed by Yeast RNA-Binding Proteins Nrd1 and Nab3 // Mol. Cell. 2006. Vol. 23, № 6.
347. Thiebaut M. et al. Transcription Termination and Nuclear Degradation of Cryptic Unstable Transcripts: A Role for the Nrd1-Nab3 Pathway in Genome Surveillance // Mol. Cell. 2006. Vol. 23, № 6.
348. Wyers F. et al. Cryptic Pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase // Cell. 2005. Vol. 121, № 5.
349. Steinmetz E.J. et al. Genome-Wide Distribution of Yeast RNA Polymerase II and Its Control by Sen1 Helicase // Mol. Cell. 2006. Vol. 24, № 5.
350. Kim K.Y., Levin D.E. Mpk1 MAPK association with the paf1 complex blocks sen1-mediated premature transcription termination // Cell. 2011. Vol. 144, № 5.
351. Nedeia E. et al. Organization and function of APT, a subcomplex of the yeast cleavage and polyadenylation factor involved in the formation of mRNA and small nucleolar RNA 3'-ends // J. Biol. Chem. 2003. Vol. 278, № 35.
352. Lidschreiber M. et al. The APT complex is involved in non-coding RNA transcription and is distinct from CPF // Nucleic Acids Res. 2018. Vol. 46, № 21. P. 11528–11538.
353. Preker P. et al. RNA exosome depletion reveals transcription upstream of active human promoters // Science (80-.). 2008. Vol. 322, № 5909.
354. Preker P. et al. PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters // Nucleic Acids Res. 2011. Vol. 39, № 16.
355. Pelechano V., Steinmetz L.M. Gene regulation by antisense transcription // Nature

- 356. Zhao L. et al. Antisense transcription and its roles in adaption to environmental stress in *E. coli* // *bioRxiv*. 2023.
- 357. Davidson L. et al. Rapid Depletion of DIS3, EXOSC10, or XRN2 Reveals the Immediate Impact of Exoribonucleolysis on Nuclear RNA Metabolism and Transcriptional Control // *Cell Rep*. 2019. Vol. 26, № 10.
- 358. Iasillo C. et al. ARS2 is a general suppressor of pervasive transcription // *Nucleic Acids Res*. 2017. Vol. 45, № 17.
- 359. Ogami K. et al. An Mtr4/ZFC3H1 complex facilitates turnover of unstable nuclear RNAs to prevent their cytoplasmic transport and global translational repression // *Genes Dev*. 2017. Vol. 31, № 12.
- 360. Elrod N.D. et al. The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes // *Mol. Cell*. 2019. Vol. 76, № 5.
- 361. Tatomer D.C. et al. The Integrator complex cleaves nascent mRNAs to attenuate transcription // *Genes Dev*. 2019. Vol. 33, № 21–22.
- 362. Lykke-Andersen S. et al. Integrator is a genome-wide attenuator of non-productive transcription // *Mol. Cell*. 2021. Vol. 81, № 3.
- 363. Stein C.B. et al. Integrator endonuclease drives promoter-proximal termination at all RNA polymerase II-transcribed loci // *Mol. Cell*. 2022. Vol. 82, № 22.
- 364. Austenaa L.M.I. et al. A first exon termination checkpoint preferentially suppresses extragenic transcription // *Nat. Struct. Mol. Biol*. 2021. Vol. 28, № 4.
- 365. Estell C. et al. Zc3h4 restricts non-coding transcription in human cells // *Elife*. 2021. Vol. 10.
- 366. Estell C. et al. A restrictor complex of ZC3H4, WDR82, and ARS2 integrates with PNUTS to control unproductive transcription // *Mol. Cell*. 2023. Vol. 83, № 13.
- 367. Rouvière J.O. et al. ARS2 instructs early transcription termination-coupled RNA decay by recruiting ZC3H4 to nascent transcripts // *Mol. Cell*. 2023. Vol. 83, № 13.
- 368. Wilson M.D., Harreman M., Svejstrup J.Q. Ubiquitylation and degradation of

- elongating RNA polymerase II: The last resort // *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. 2013. Vol. 1829, № 1.
369. Gullerova M., Proudfoot N.J. Transcriptional interference and gene orientation in yeast: Noncoding RNA connections // *Cold Spring Harb. Symp. Quant. Biol.* 2010. Vol. 75.
370. Dhir A. et al. Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs // *Nat. Struct. Mol. Biol.* 2015. Vol. 22, № 4.
371. Lemay J.F. et al. The RNA exosome promotes transcription termination of backtracked RNA polymerase II // *Nat. Struct. Mol. Biol.* 2014. Vol. 21, № 10.
372. Reynolds D. et al. Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in *Leishmania major* and *Trypanosoma brucei* // *Nucleic Acids Res.* 2014. Vol. 42, № 15.
373. Soldatova I. V. et al. A Novel Mechanism for Transcription Termination in the mod(mdg4) Locus of *Drosophila melanogaster* // *Biology (Basel)*. 2024. Vol. 13, № 12.
374. Tikhonov M. et al. Conserved sequences in the *Drosophila* mod(mdg4) intron promote poly(A)-independent transcription termination and trans-splicing // *Nucleic Acids Res.* 2018. Vol. 46, № 20.
375. Vilborg A. et al. Widespread Inducible Transcription Downstream of Human Genes // *Mol. Cell*. 2015. Vol. 59, № 3.
376. Nemeroff M.E. et al. Influenza virus NS1 protein interacts with the cellular 30 kDa subunit of CPSF and inhibits 3' end formation of cellular pre-mRNAs // *Mol. Cell*. 1998. Vol. 1, № 7.
377. Rutkowski A.J. et al. Widespread disruption of host transcription termination in HSV-1 infection // *Nat. Commun.* 2015. Vol. 6.
378. Grosso A.R. et al. Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma // *Elife*. 2015. Vol. 4, № NOVEMBER2015.

379. Wang X. et al. Herpes simplex virus blocks host transcription termination via the bimodal activities of ICP27 // *Nat. Commun.* 2020. Vol. 11, № 1.
380. Tang S., Patel A., Krause P.R. Herpes simplex virus ICP27 regulates alternative pre-mRNA polyadenylation and splicing in a sequence-dependent manner // *Proc. Natl. Acad. Sci. U. S. A.* 2016. Vol. 113, № 43.
381. Wang X. et al. Mechanism and consequences of herpes simplex virus 1-mediated regulation of host mRNA alternative polyadenylation // *PLoS Genet.* 2021. Vol. 17, № 3.
382. Bauer D.L.V. et al. Influenza Virus Mounts a Two-Pronged Attack on Host RNA Polymerase II Transcription // *Cell Rep.* 2018. Vol. 23, № 7.
383. Davidson L., Muniz L., West S. 3' end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells // *Genes Dev.* 2014. Vol. 28, № 4.
384. Das K. et al. Structural basis for suppression of a host antiviral response by influenza A virus // *Proc. Natl. Acad. Sci. U. S. A.* 2008. Vol. 105, № 35.
385. Vo J.N. et al. The Landscape of Circular RNA in Cancer // *Cell.* 2019. Vol. 176, № 4.
386. Vidal A.F. Read-through circular RNAs reveal the plasticity of RNA processing mechanisms in human cells // *RNA Biol.* 2020. Vol. 17, № 12.
387. Xing H. et al. HSF1 Modulation of Hsp70 mRNA Polyadenylation via Interaction with Symplekin // *J. Biol. Chem.* 2004. Vol. 279, № 11.
388. Mohn F. et al. The Rhino-Deadlock-Cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in *Drosophila* // *Cell.* 2014. Vol. 157, № 6.
389. Zhang Z. et al. The HP1 homolog Rhino anchors a nuclear complex that suppresses piRNA precursor splicing // *Cell.* 2014. Vol. 157, № 6.
390. Le Thomas A. et al. Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor

- processing // *Genes Dev.* 2014. Vol. 28, № 15.
391. Dieci G. et al. Transcription reinitiation by RNA polymerase III // *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. 2013. Vol. 1829, № 3–4.
392. Arimbasseri A.G., Rijal K., Maraia R.J. Transcription termination by the eukaryotic RNA polymerase III // *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. 2013. Vol. 1829, № 3–4.
393. Orioli A. et al. Widespread occurrence of non-canonical transcription termination by human RNA polymerase III // *Nucleic Acids Res.* 2011. Vol. 39, № 13.
394. Braglia P., Percudani R., Dieci G. Sequence context effects on oligo(dT) termination signal recognition by *Saccharomyces cerevisiae* RNA polymerase III // *J. Biol. Chem.* 2005. Vol. 280, № 20.
395. Gunnery S., Ma Y., Mathews M.B. Termination sequence requirements vary among genes transcribed by RNA polymerase III // *J. Mol. Biol.* 1999. Vol. 286, № 3.
396. Nielsen S., Yuzenkova Y., Zenkin N. Mechanism of eukaryotic RNA polymerase III transcription termination // *Science* (80-.). 2013. Vol. 340, № 6140.
397. Patwardhan R.P. et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis // *Nat. Biotechnol.* 2009. Vol. 27, № 12.
398. Kinney J.B. et al. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence // *Proc. Natl. Acad. Sci. U. S. A.* 2010. Vol. 107, № 20.
399. Kosuri S. et al. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli* // *Proc. Natl. Acad. Sci. U. S. A.* 2013. Vol. 110, № 34.
400. Sharon E. et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters // *Nat. Biotechnol.* 2012. Vol. 30, № 6.
401. Dvir S. et al. Deciphering the rules by which 5'-UTR sequences affect protein

- expression in yeast // Proc. Natl. Acad. Sci. U. S. A. 2013. Vol. 110, № 30.
402. Melnikov A. et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay // Nat. Biotechnol. 2012. Vol. 30, № 3.
403. Oikonomou P., Goodarzi H., Tavazoie S. Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts // Cell Rep. The Authors, 2014. Vol. 7, № 1. P. 281–292.
404. Zhao W. et al. Massively parallel functional annotation of 3' untranslated regions // Nat. Biotechnol. Nature Publishing Group, 2014. Vol. 32, № 4. P. 387–391.
405. Patwardhan R.P. et al. Massively parallel functional dissection of mammalian enhancers in vivo // Nat. Biotechnol. 2012. Vol. 30, № 3.
406. La Fleur A., Shi Y., Seelig G. Decoding biology with massively parallel reporter assays and machine learning // Genes Dev. 2024. Vol. 38, № 17–20. P. 843–865.
407. Kwasnieski J.C. et al. Complex effects of nucleotide variants in a mammalian cis-regulatory element // Proc. Natl. Acad. Sci. U. S. A. 2012. Vol. 109, № 47.
408. Mogno I., Kwasnieski J.C., Cohen B.A. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants // Genome Res. 2013. Vol. 23, № 11.
409. Van Arensbergen J. et al. Genome-wide mapping of autonomous promoter activity in human cells // Nat. Biotechnol. 2017. Vol. 35, № 2.
410. Arnold C.D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq // Science (80-.). 2013. Vol. 339, № 6123.
411. Kheradpour P. et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay // Genome Res. 2013. Vol. 23, № 5.
412. Ke S. et al. Quantitative evaluation of all hexamers as exonic splicing elements // Genome Res. 2011. Vol. 21, № 8.
413. Noderer W.L. et al. Quantitative analysis of mammalian translation initiation sites

- by FACS - seq // Mol. Syst. Biol. 2014. Vol. 10, № 8.
414. Cuperus J.T. et al. Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences // Genome Res. 2017. Vol. 27, № 12.
 415. Cambray G., Guimaraes J.C., Arkin A.P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in escherichia coli // Nature Biotechnology. 2018. Vol. 36, № 10.
 416. Zhao W. et al. Massively parallel functional annotation of 3' untranslated regions // Nat. Biotechnol. 2014. Vol. 32, № 4. P. 387–391.
 417. Rosenberg A.B. et al. Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences // Cell. 2015. Vol. 163, № 3.
 418. Paggi J. et al. Predicting transcriptional regulatory activities with deep convolutional networks // bioRxiv. 2017.
 419. Movva R. et al. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays // PLoS One. 2019. Vol. 14, № 6.
 420. Sample P.J. et al. Human 5' UTR design and variant effect prediction from a massively parallel translation assay // Nat. Biotechnol. 2019. Vol. 37, № 7.
 421. Linder J. et al. Deciphering the impact of genetic variation on human polyadenylation using APARENT2 // Genome Biol. BioMed Central, 2022. Vol. 23, № 1. P. 1–33.
 422. de Boer C.G., Taipale J. Hold out the genome: a roadmap to solving the cis-regulatory code // Nature. Springer US, 2024. Vol. 625, № 7993. P. 41–50.
 423. de Boer C.G. et al. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters // Nat. Biotechnol. 2020. Vol. 38, № 1. P. 56–65.
 424. White M.A. et al. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks // Proc. Natl. Acad. Sci. U. S. A. 2013. Vol. 110, № 29.
 425. Rabani M. et al. A Massively Parallel Reporter Assay of 3' UTR Sequences

- Identifies In Vivo Rules for mRNA Degradation // Mol. Cell. Elsevier Inc., 2017. Vol. 68, № 6. P. 1083-1094.e5.
426. Verfaillie A. et al. Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic // Genome Res. 2016. Vol. 26, № 7.
427. Cottrell K.A. et al. PTRE-seq reveals mechanism and interactions of RNA binding proteins and miRNAs // Nat. Commun. 2018. Vol. 9, № 1.
428. de Almeida B.P. et al. Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo // Nature. 2024. Vol. 626, № 7997.
429. Taskiran I.I. et al. Cell-type-directed design of synthetic enhancers // Nature. 2024. Vol. 626, № 7997.
430. Mikl M. et al. Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries // Nat. Commun. 2019. Vol. 10, № 1.
431. May G.E. et al. Unraveling the influences of sequence and position on yeast uORF activity using massively parallel reporter systems and machine learning // Elife. 2023. Vol. 12.
432. Chong R. et al. A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions // Mol. Cell. 2019. Vol. 73, № 1.
433. Peterman N., Levine E. Sort-seq under the hood: Implications of design choices on large-scale characterization of sequence-function relations // BMC Genomics. 2016. Vol. 17, № 1.
434. Kim I. et al. High-throughput analysis of in vivo protein stability // Mol. Cell. Proteomics. 2013. Vol. 12, № 11.
435. Liachko I. et al. High-resolution mapping, characterization, and optimization of autonomously replicating sequences in yeast // Genome Res. 2013. Vol. 23, № 4.
436. Wissink E.M., Fogarty E.A., Grimson A. High-throughput discovery of post-transcriptional cis-regulatory elements // BMC Genomics. BMC Genomics, 2016. Vol. 17, № 1. P. 1–14.

437. Niederer R.O. et al. Direct analysis of ribosome targeting illuminates thousand-fold regulation of translation initiation // *Cell Syst.* 2022. Vol. 13, № 3.
438. Love M.I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 // *Genome Biol.* 2014. Vol. 15, № 12.
439. Chen Y. et al. edgeR 4.0: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets // *bioRxiv.* 2024.
440. Ashuach T. et al. MPRAalyze: Statistical framework for massively parallel reporter assays // *Genome Biol.* 2019. Vol. 20, № 1.
441. Myint L. et al. Linear models enable powerful differential activity analysis in massively parallel reporter assays // *BMC Genomics.* 2019. Vol. 20, № 1.
442. Gordon M.G. et al. lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements // *Nat. Protoc.* 2020. Vol. 15, № 8.
443. Shalem O. et al. Systematic Dissection of the Sequence Determinants of Gene 3' End Mediated Expression Control // *PLoS Genetics.* 2015. Vol. 11, № 4.
444. Shen S.Q. et al. Massively parallel cis-regulatory analysis in the mammalian central nervous system // *Genome Res.* 2016. Vol. 26, № 2.
445. Leppek K. et al. Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics // *Nat. Commun.* 2022. Vol. 13, № 1.
446. Chiang H.L. et al. Mechanism and modeling of human disease-associated near-exon intronic variants that perturb RNA splicing // *Nat. Struct. Mol. Biol.* 2022. Vol. 29, № 11.
447. Soemedi R. et al. Pathogenic variants that alter protein code often disrupt splicing // *Nat. Genet.* 2017. Vol. 49, № 6.
448. Mikl M. et al. A massively parallel reporter assay reveals focused and broadly encoded RNA localization signals in neurons // *Nucleic Acids Res.* 2022. Vol. 50, № 18.
449. Cao J. et al. High-throughput 5' UTR engineering for enhanced protein production

- in non-viral gene therapies // Nat. Commun. 2021. Vol. 12, № 1.
450. Cheng J.K. et al. Design and Evaluation of Synthetic Terminators for Regulating Mammalian Cell Transgene Expression // ACS Synth. Biol. 2019. Vol. 8, № 6. P. 1263–1275.
 451. Penzar D. et al. LegNet: A best-in-class deep learning model for short DNA regulatory regions // Bioinformatics. 2023. Vol. 39, № 8.
 452. Quang D., Xie X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences // Nucleic Acids Res. 2016. Vol. 44, № 11.
 453. Angermueller C. et al. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning // Genome Biol. 2017. Vol. 18, № 1.
 454. Rives A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences // Proc. Natl. Acad. Sci. U. S. A. 2021. Vol. 118, № 15.
 455. Lin Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model // Science (80-.). 2023. Vol. 379, № 6637.
 456. Ji Y. et al. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome // Bioinformatics. 2021. Vol. 37, № 15.
 457. Consens M.E. et al. To Transformers and Beyond: Large Language Models for the Genome. 2023.
 458. Karollus A. et al. Species-aware DNA language models capture regulatory elements and their evolution // Genome Biol. 2024. Vol. 25, № 1.
 459. Chu Y. et al. A 5' UTR language model for decoding untranslated regions of mRNA and function predictions // Nat. Mach. Intell. 2024. Vol. 6, № 4.
 460. Chen K. et al. Self-supervised learning on millions of primary RNA sequences from 72 vertebrates improves sequence-based RNA splicing prediction // Brief. Bioinform. 2024. Vol. 25, № 3.

461. Biswas S. et al. Low-N protein engineering with data-efficient deep learning // Nat. Methods. 2021. Vol. 18, № 4.
462. de Almeida B.P. et al. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers // Nat. Genet. 2022. Vol. 54, № 5.
463. Gosai S. et al. Machine-guided design of synthetic cell type-specific cis-regulatory elements // bioRxiv. 2023.
464. Castillo-Hair S. et al. Optimizing 5'UTRs for mRNA-delivered gene editing using deep learning // Nat. Commun. 2024. Vol. 15, № 1.
465. Killoran N. et al. Generating and designing DNA with deep generative models // arXiv. 2017.
466. Brookes D.H., Listgarten J. Design by adaptive sampling // arXiv. 2018.
467. Costello Z., Martin H.G. How to Hallucinate Functional Proteins // arXiv. 2019.
468. Gupta A., Zou J. Feedback GAN for DNA optimizes protein functions // Nat. Mach. Intell. 2019. Vol. 1, № 2.
469. Linder J. et al. A Generative Neural Network for Maximizing Fitness and Diversity of Synthetic DNA and Protein Sequences // Cell Syst. 2020. Vol. 11, № 1.
470. Repecka D. et al. Expanding functional protein sequence spaces using generative adversarial networks // Nat. Mach. Intell. 2021. Vol. 3, № 4.
471. Shin J.E. et al. Protein design and variant prediction using autoregressive generative models // Nat. Commun. 2021. Vol. 12, № 1.
472. Zrimec J. et al. Controlling gene expression with deep generative design of regulatory DNA // Nat. Commun. 2022. Vol. 13, № 1.
473. Uehara M. et al. Bridging Model-Based Optimization and Generative Modeling via Conservative Fine-Tuning of Diffusion Models // arXiv. 2024. P. 1–30.
474. Lal A. et al. regLM: Designing Realistic Regulatory DNA with Autoregressive Language Models // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2024.

Vol. 14758 LNCS.

475. Avdeyev P. et al. Dirichlet Diffusion Score Model for Biological Sequence Generation // Proceedings of Machine Learning Research. 2023. Vol. 202.
476. DaSilva L.F. et al. DNA-Diffusion: Leveraging Generative Models for Controlling Chromatin Accessibility and Gene Expression via Synthetic Regulatory Elements // bioRxiv. 2024.
477. Sarkar A. et al. Designing DNA With Tunable Regulatory Activity Using Discrete Diffusion // bioRxiv. 2024.
478. Stark H. et al. Dirichlet Flow Matching with Applications to DNA Sequence Design // Proc. Mach. Learn. Res. 2024. Vol. 235. P. 46495–46513.
479. Yin C. et al. Iterative deep learning-design of human enhancers exploits condensed sequence grammar to achieve cell type-specificity // bioRxiv. 2024.
480. Friedman R.Z. et al. Active learning of enhancer and silencer regulatory grammar in photoreceptors // bioRxiv. 2023.
481. Landrum M.J. et al. ClinVar: Improving access to variant interpretations and supporting evidence // Nucleic Acids Res. 2018. Vol. 46, № D1.
482. Cui Y. et al. 3'aQTL-atlas: An atlas of 3'UTR alternative polyadenylation quantitative trait loci across human normal tissues // Nucleic Acids Res. 2022. Vol. 50, № D1.
483. Li L. et al. An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability // Nat. Genet. 2021. Vol. 53, № 7.
484. Kowalski M.H. et al. Multiplexed single-cell characterization of alternative polyadenylation regulators // Cell. Cell Press, 2024. Vol. 187, № 16. P. 4408-4425.e23.
485. Zhu B. et al. The human PAF complex coordinates transcription with events downstream of RNA synthesis // Genes Dev. 2005. Vol. 19, № 14.
486. Katahira J. mRNA export and the TREX complex // Biochimica et Biophysica

487. Gil A., Proudfoot N.J. Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit β -globin mRNA 3' end formation // Cell. 1987. Vol. 49, № 3. P. 399–406.
488. Gil A., Proudfoot N.J. A sequence downstream of AAUAAA is required for rabbit β -globin mRNA 3'-end formation // Nature. 1984. Vol. 312, № 5993. P. 473–474.
489. Mclauchlan J. et al. The consensus sequence YGTGTTY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini // Nucleic Acids Res. 1985. Vol. 13, № 4.
490. Birnstiel M.L.A., Busslinger M., Strub K. Transcription termination and 3' processing: the end is in site! // Cell. 1985. Vol. 41, № 2.
491. Taya Y. et al. Cloning and structure of the human immune interferon-gamma chromosomal gene. // EMBO J. 1982. Vol. 1, № 8.
492. Berget S.M. Are U4 small nuclear ribonucleoproteins involved in polyadenylation? // Nature. 1984. Vol. 309, № 5964.
493. McDevitt M.A. et al. Requirement of a downstream sequence for generation of a poly(A) addition site // Cell. 1984. Vol. 37, № 3.
494. Cole C.N., Stacy T.P. Identification of sequences in the herpes simplex virus thymidine kinase gene required for efficient processing and polyadenylation. // Mol. Cell. Biol. 1985. Vol. 5, № 8.
495. Conway L., Wickens M. A sequence downstream of A-A-U-A-A-A is required for formation of simian virus 40 late mRNA 3' termini in frog oocytes // Proc. Natl. Acad. Sci. U. S. A. 1985. Vol. 82, № 12.
496. Wilusz J., Shenk T. A Uridylate Tract Mediates Efficient Heterogeneous Nuclear Ribonucleoprotein C Protein-RNA Cross-Linking and Functionally Substitutes for the Downstream Element of the Polyadenylation Signal // Mol. Cell. Biol. 1990. Vol. 10, № 12. P. 6397–6407.
497. Sadofsky M. et al. Identification of a sequence element on the 3' side of AAUAAA

- which is necessary for simian virus 40 late mRNA 3'-end processing. // *Mol. Cell. Biol.* 1985. Vol. 5, № 10.
498. Sittler A., Gallinaro H., Jacob M. Upstream and downstream c/s-acting elements for cleavage at the L4 polyadenylation site of adenovirus-2 // *Nucleic Acids Res.* 1994. Vol. 22, № 2.
 499. Hart R.P. et al. Definition of Essential Sequences and Functional Equivalence of Elements Downstream of the Adenovirus E2A and the Early Simian Virus 40 Polyadenylation Sites // *Mol. Cell. Biol.* 1985. Vol. 5, № 11. P. 2975–2983.
 500. McDevitt M.A. et al. Sequences capable of restoring poly(A) site function define two distinct downstream elements. // *EMBO J.* 1986. Vol. 5, № 11. P. 2907–2913.
 501. Chou Z. fang, Chen F., Wilusz J. Sequence and position requirements for uridylate-rich downstream elements of polyadenylation signals // *Nucleic Acids Res.* 1994. Vol. 22, № 13. P. 2525–2531.
 502. Zarkower D., Wickens M. A functionally redundant downstream sequence in SV40 late pre-mRNA is required for mRNA 3'-end formation and for assembly of a precleavage complex in vitro // *J. Biol. Chem.* © 1988 ASBMB. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology., 1988. Vol. 263, № 12. P. 5780–5788.
 503. Goodwin E.C., Rottman F.M. The 3'-flanking sequence of the bovine growth hormone gene contains novel elements required for efficient and accurate polyadenylation // *J. Biol. Chem.* 1992. Vol. 267, № 23. P. 16330–16334.
 504. Zarudnaya M.I. et al. Downstream elements of mammalian pre-mRNA polyadenylation signals: Primary, secondary and higher-order structures // *Nucleic Acids Res.* 2003. Vol. 31, № 5. P. 1375–1386.
 505. Letiagina A.E. et al. 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.
 506. Lorenz R. et al. ViennaRNA Package 2.0 // *Algorithms Mol. Biol.* 2011. Vol. 6, № 1.

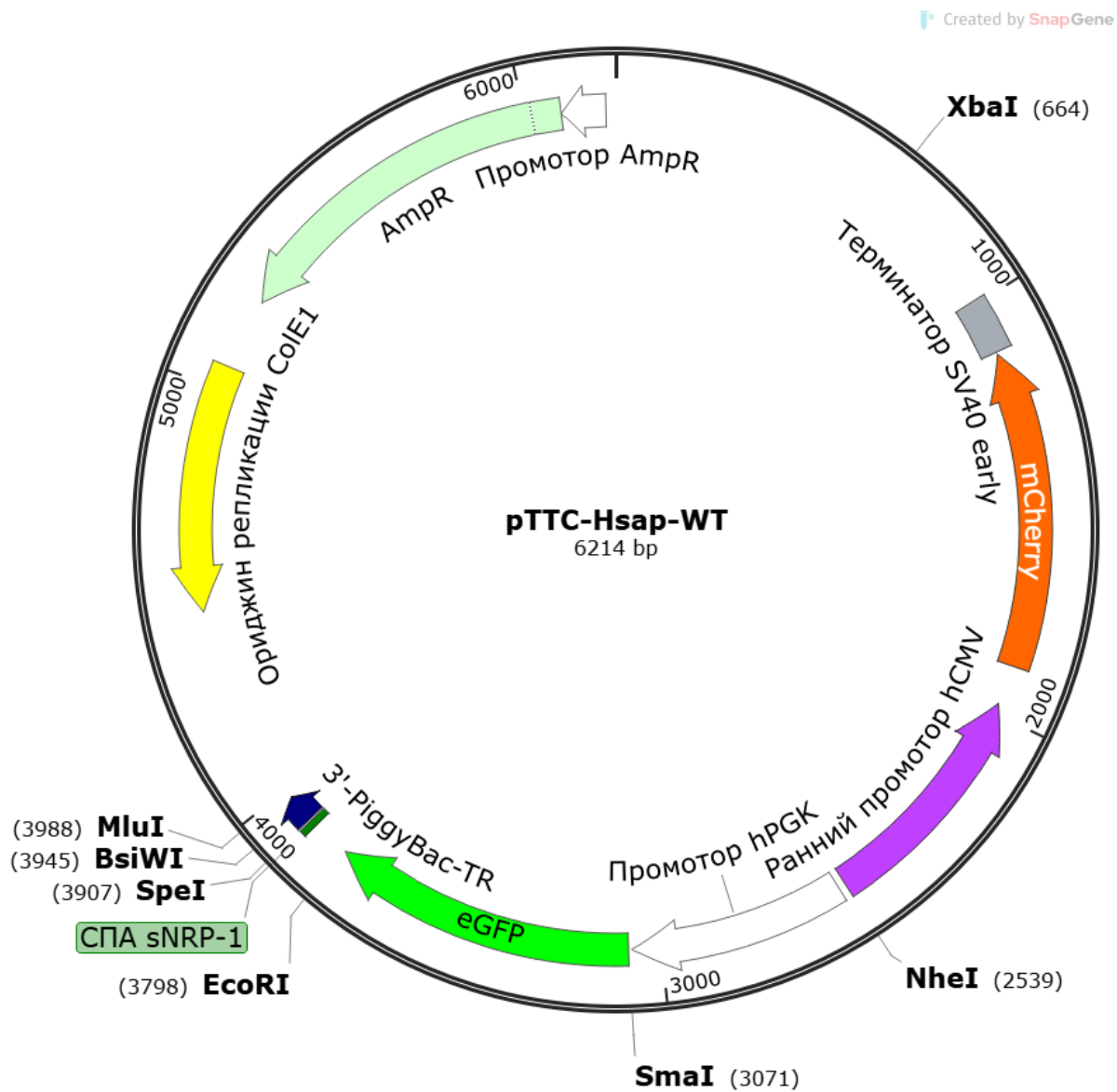
507. Akiba T. et al. Optuna: A Next-generation Hyperparameter Optimization Framework // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019.
508. Komura R. et al. High-throughput evaluation of T7 promoter variants using biased randomization and DNA barcoding // PLoS One. 2018. Vol. 13, № 5.
509. Tewhey R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay // Cell. 2016. Vol. 165, № 6.
510. Shao K. et al. Emulsion PCR: A high efficient way of PCR amplification of random DNA libraries in aptamer selection // PLoS One. 2011. Vol. 6, № 9.
511. Meyerhans A., Vartanian J.P., Wain-Hobson S. DNA recombination during PCR // Nucleic Acids Res. 1990. Vol. 18, № 7.
512. Qiu X. et al. Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning // Appl. Environ. Microbiol. 2001. Vol. 67, № 2.
513. Wang G.C.Y., Wang Y. The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species // Microbiology. 1996. Vol. 142, № 5.
514. Wang G.C.Y., Wang Y. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes // Appl. Environ. Microbiol. 1997. Vol. 63, № 12.
515. Acinas S.G. et al. PCR-induced sequence artifacts and bias: Insights from comparison of two 16s rRNA clone libraries constructed from the same sample // Appl. Environ. Microbiol. 2005. Vol. 71, № 12.
516. Bjørnsgaard Aas A., Davey M.L., Kauserud H. ITS all right mama: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities // Mol. Ecol. Resour. 2017. Vol. 17, № 4.
517. Cronn R. et al. PCR-mediated recombination in amplification products derived from polyploid cotton // Theor. Appl. Genet. 2002. Vol. 104, № 2–3.

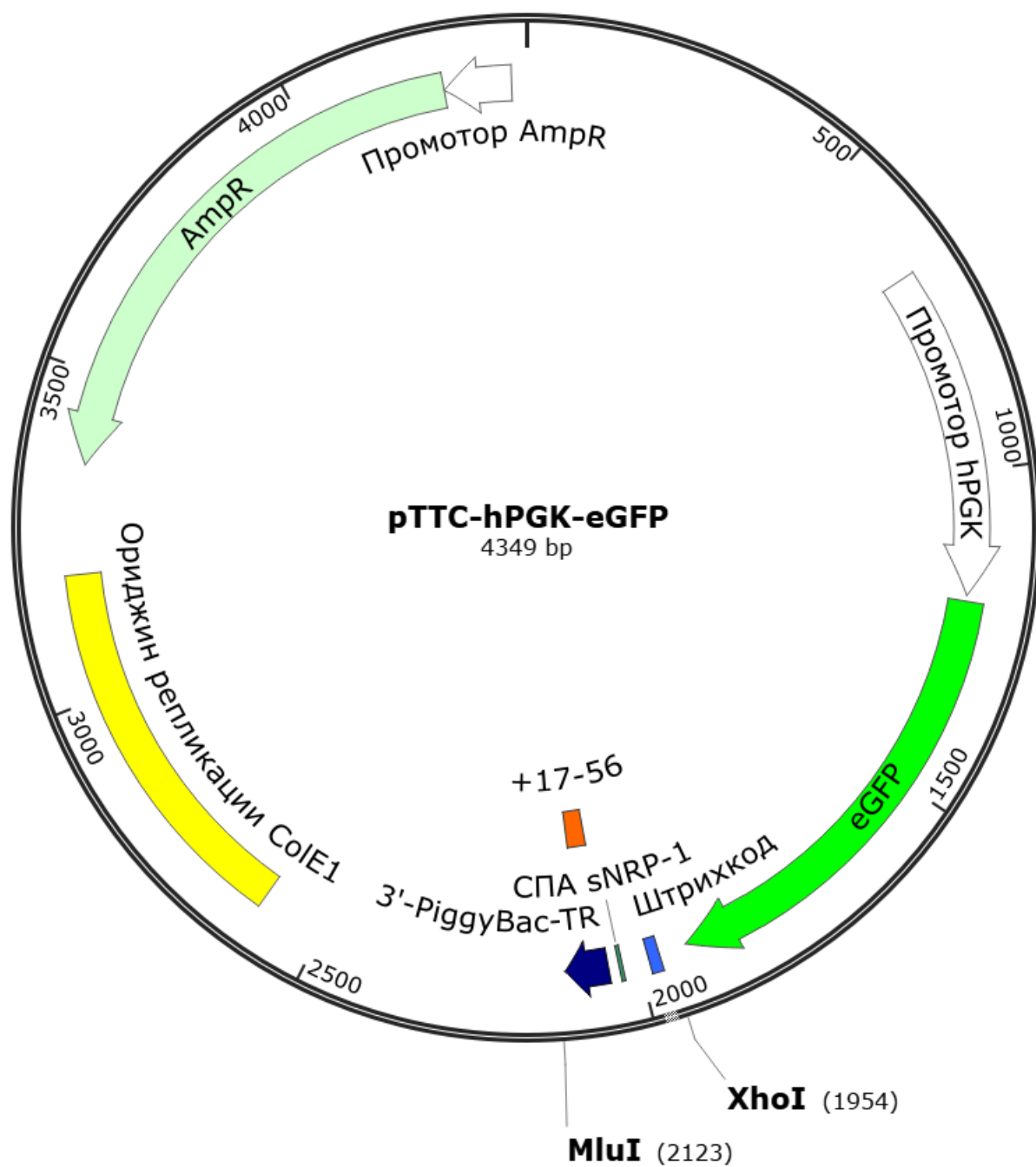
518. Potapov V., Ong J.L. Examining sources of error in PCR by single-molecule sequencing // PLoS One. 2017. Vol. 12, № 1.
519. Omelina E.S. et al. Slight Variations in the Sequence Downstream of the Polyadenylation Signal Significantly Increase Transgene Expression in HEK293T and CHO Cells // Int. J. Mol. Sci. 2022. Vol. 23, № 24.
520. Bradley R.D., Hillis D.M. Recombinant DNA sequences generated by PCR amplification. // Molecular biology and evolution. 1997. Vol. 14, № 5.
521. Klug J., Wolf M., Beato M. Creating chimeric molecules by PCR directed homologous DNA recombination // Nucleic Acids Res. 1991. Vol. 19, № 10.
522. Omelina E.S. et al. Optimized PCR conditions minimizing the formation of chimeric DNA molecules from MPRA plasmid libraries // BMC Genomics. 2019. Vol. 20, № Suppl 7. P. 1–10.
523. Lahr D.J.G., Katz L.A. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase // Biotechniques. 2009. Vol. 47, № 4.
524. Liu J. et al. Extensive recombination due to heteroduplexes generates large amounts of artificial gene fragments during pcr // PLoS One. 2014. Vol. 9, № 9.
525. Boers S.A., Hays J.P., Jansen R. Micelle PCR reduces chimera formation in 16S rRNA profiling of complex microbial DNA mixtures // Sci. Rep. 2015. Vol. 5.
526. Williams R. et al. Amplification of complex gene libraries by emulsion PCR // Nat. Methods. 2006. Vol. 3, № 7.
527. O'Shea J.P. et al. PLogo: A probabilistic approach to visualizing sequence motifs // Nat. Methods. 2013. Vol. 10, № 12. P. 1211–1212.
528. Gimmi E.R., Reff M.E., Deckman I.C. Alterations in the pre-mRNA topology of the bovine growth hormone polyadenylation region decrease poly(A) site efficiency // Nucleic Acids Res. 1989. Vol. 17, № 17.
529. Herrmann C.J. et al. PolyASite 2.0: A consolidated atlas of polyadenylation sites from 3' end sequencing // Nucleic Acids Res. 2020. Vol. 48, № D1.

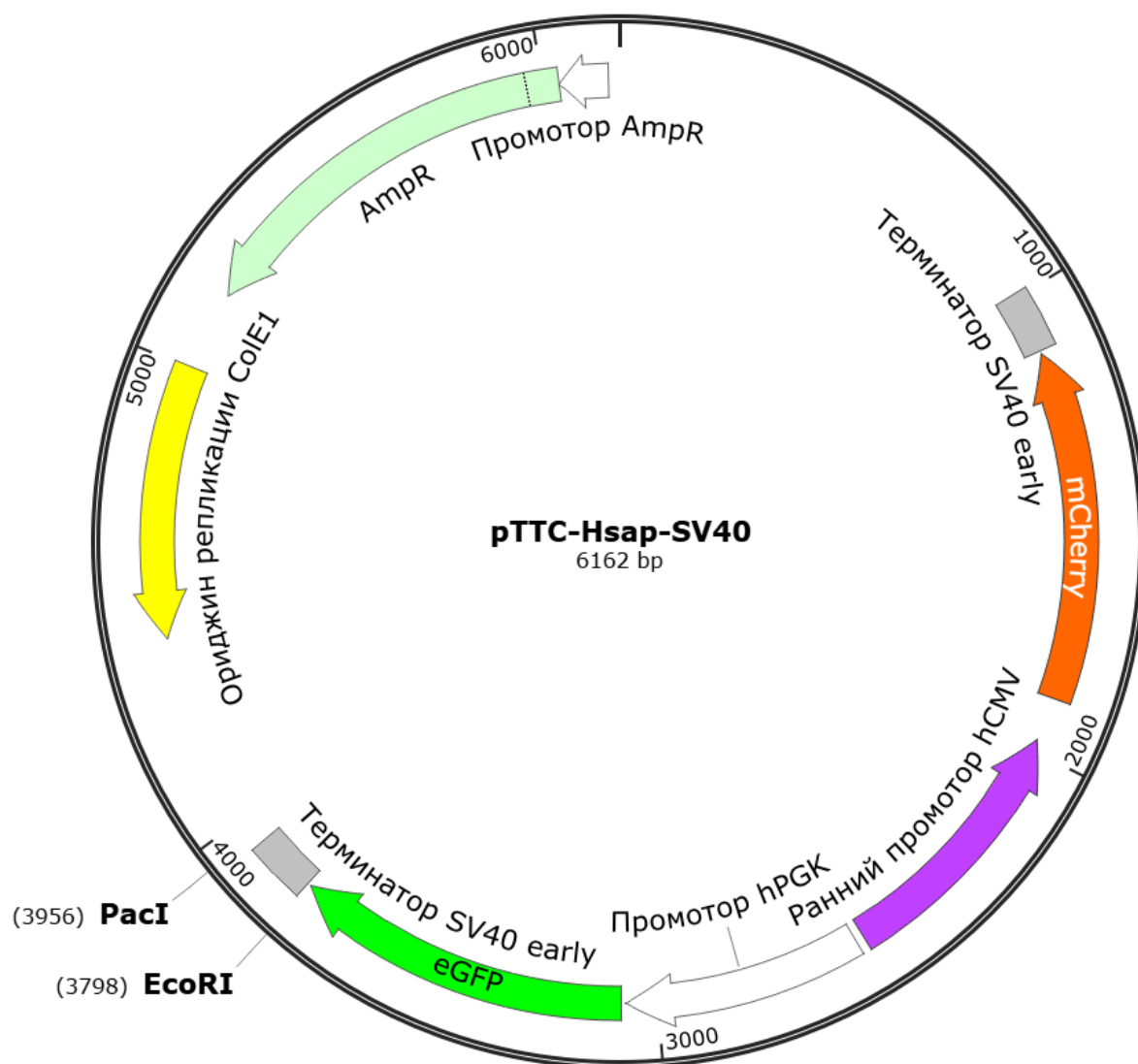
530. Kerimov N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci // Nat. Genet. Springer US, 2021. Vol. 53, № 9. P. 1290–1299.
531. Ashfield R. et al. MAZ-dependent termination between closely spaced human complement genes // EMBO J. 1994. Vol. 13, № 23. P. 5656–5667.
532. Yonaha M., Proudfoot N.J. Specific transcriptional pausing activates polyadenylation in a coupled in vitro system // Mol. Cell. 1999. Vol. 3, № 5. P. 593–600.
533. Levitt N. et al. Definition of an efficient synthetic poly(A) site. // Genes Dev. 1989. Vol. 3, № 7. P. 1019–1025.
534. Deng Z. et al. Useful bicistronic reporter system for studying Poly(A) site-defining cis elements and regulation of alternative polyadenylation // Int. J. Mol. Sci. 2018. Vol. 19, № 1.
535. Davidson L. et al. DNA-directed termination of mammalian RNA polymerase II // Genes Dev. 2024. Vol. 38, № 21–24. P. 998–1019.

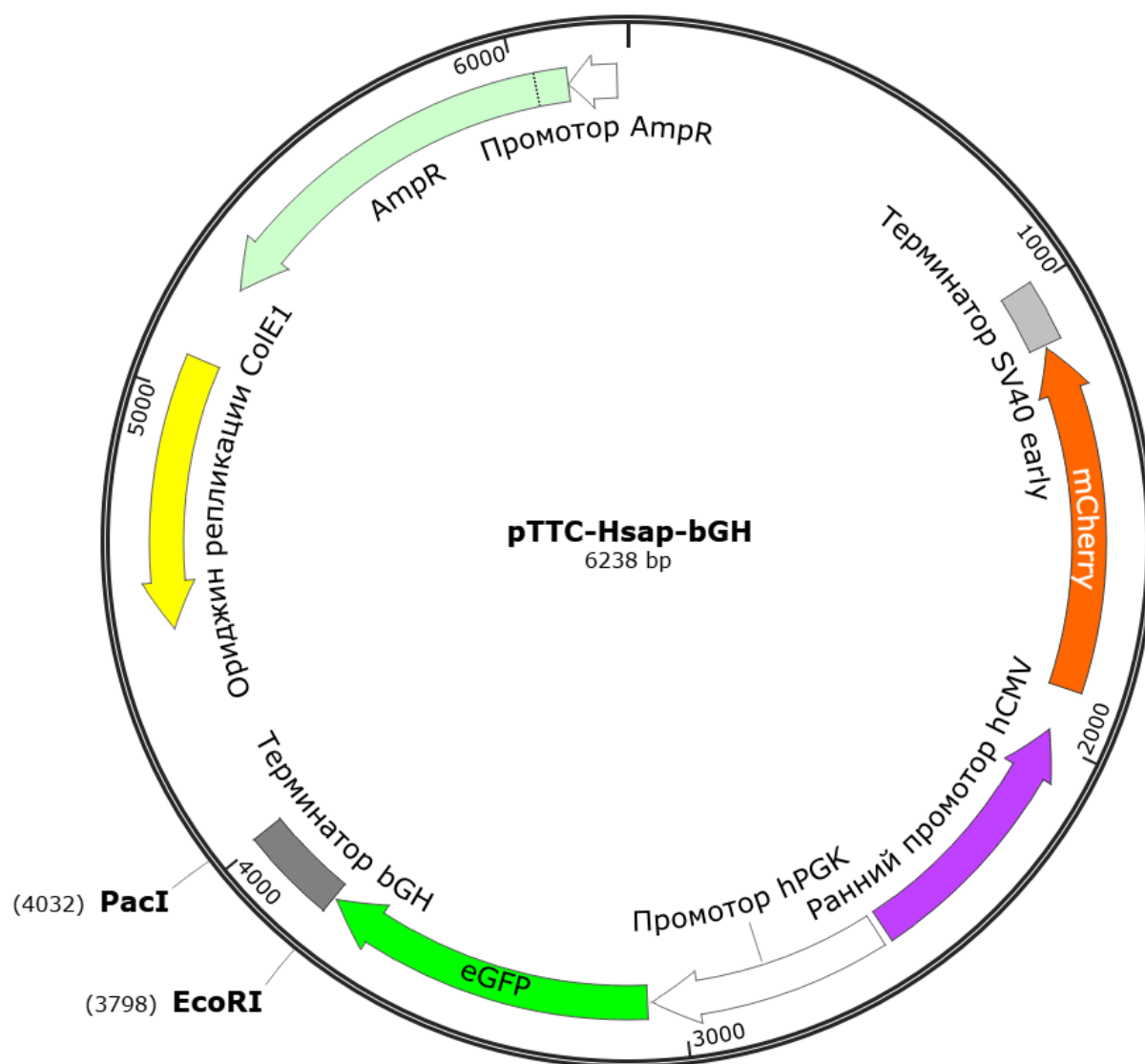
Приложение 1

Карты использованных в работе плазмид









Приложение 2

Список использованных праймеров

№	Название праймера	Последовательность, 5'→3'	Применение
1	pTTC-hPGK-eGFP-BC-F	GGTGACACTCGAGGAT CGAGNNNNNNNNNNNN NNNNNNGAGTTGTGGC CGGC	Амплификация встройки для конструирования плазмидных библиотек
2	pTTC-hPGK-eGFP-17-24	AGATAATCATGCGTAA AATTNNNNNNNNNACTA GTCATTTTCGTATTTTATT TAC	Амплификация встройки для конструирования плазмидной библиотеки 17-24
3	pTTC-hPGK-eGFP-21-28	TTAAAGATAATCATGCG TAANNNNNNNNNCATGA CTAGTCATTTTCG	Амплификация встройки для конструирования плазмидной библиотеки 21-28
4	pTTC-hPGK-eGFP-25-32-R	TACGTTAAGATAATCA TGCNNNNNNNNNGACGC ATGACTAGTCATTTTC	Амплификация встройки для конструирования плазмидной библиотеки 25-32
5	pTTC-hPGK-eGFP-29-36	ACGTACGTTAAGATA ATCNNNNNNNNNAATTG ACGCATGACTAGTC	Амплификация встройки для конструирования плазмидной библиотеки 29-36
6	pTTC-hPGK-eGFP-33-40	TGTGACGTACGTTAAG ATNNNNNNNNNGTAAAA TTGACGCATGACTAG	Амплификация встройки для конструирования плазмидной библиотеки 33-40

7	pTTC- hPGK- eGFP-37-44	ATATTGTGACGTACGTT AANNNNNNNNATGCGT AAAATTGACGCATG	Амплификация встройки для конструирования плазмидной библиотеки 37-44
8	pTTC- hPGK- eGFP-41-48	AATCATATTGTGACGTA CGNNNNNNNNNAATCAT GCGTAAAATTGAC	Амплификация встройки для конструирования плазмидной библиотеки 41-48
9	pTTC- hPGK- eGFP-45-52	AGATAATCATATTGTGA CGNNNNNNNNNAGATAA TCATGCGTAAAATTG	Амплификация встройки для конструирования плазмидной библиотеки 45-52
10	pTTC- hPGK- eGFP-49-56	AGAAAGATAATCATATT GTNNNNNNNNNTTAAAG ATAATCATGCG	Амплификация встройки для конструирования плазмидной библиотеки 49-56
11	pTTC- hPGK-GA2- F	AATTTTACGCATGATTA TCTTTAACGTACGTCAC	Амплификация вектора для конструирования плазмидных библиотек 17-24 и 21-28
12	pTTC- hPGK- eGFP-GA-F	GCATGATTATCTTTAAC GTACGTCACAATATGAT TATC	Амплификация вектора для конструирования плазмидных библиотек 25-32, 29-36, 33-40, 37-44 и 41-48
13	pTTC- hPGK-GA3- F	CGTCACAATATGATTAT CTTTCTAGGGTTAATCT AGTATAC	Амплификация вектора для конструирования плазмидных библиотек 45-52 и 49-56
14	pTTC- hPGK- eGFP-GA-R	CTCGATCCTCGAGTGTC ACCTAAATC	Амплификация вектора для конструирования плазмидных библиотек
15	wt-BCI-for	GACACTCGAGGATCGA GTTCCAAGTGCAGGTTA	Амплификация встройки для конструирования контрольной плазмиды I

		GCGGAGTTGTGGCCG GCCCTTG	
16	wt-BCII-for	GACACTCGAGGATCGA GTGTGTACGGCTTGCTC TCAAGAGTTGTGGCCG CCCTTG	Амплификация встройки для конструирования контрольной плазмиды II
17	deltaC- BCIII-for	GACACTCGAGGATCGA GGAGCCCGGATCCACTC CAAGGAGTTGTGGCCG GCCCTTG	Амплификация встройки для конструирования контрольной плазмиды III
18	deltaC- BCIV-for	GACACTCGAGGATCGA GTGTCACGTCAGCTAAC CCACGAGTTGTGGCCG CCCTTG	Амплификация встройки для конструирования контрольной плазмиды IV
19	wt/deltaC- BC-rev	CGCATACGCGTATACTA GATTAACC	Амплификация встройки для конструирования контрольных плазмид, мутация сайта SpeI в промоторе hPGK
20	Libr_A1_for	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAGT TCGGAGTGACACTCGA GGATCGAG	Получение образца картирования для МПРА
21	Libr_A2_for	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG ACTCATTTGACACTCGA GGATCGAG	Получение образца картирования для МПРА
22	Libr_A3_for	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG	Получение образца картирования для МПРА

		GGGATCCG GACACTCG AGGATCGAG	
23	Libr_A4_for	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAGT CAAGCAAGACACTCGA GGATCGAG	Получение образца картирования для МПРА
24	Libr_A5_for	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG CAAGATAAGACACTCG AGGATCGAG	Получение образца картирования для МПРА
25	Libr_A6_for	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG GGACAACG GACACTCG AGGATCGAG	Получение образца картирования для МПРА
26	Libr_A7_for	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG AGCGAGCTGACACTCG AGGATCGAG	Получение образца картирования для МПРА
27	Libr_A8_for	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG CTGCACGTGACACTCGA GGATCGAG	Получение образца картирования для МПРА
28	Libr_A9_for	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG GCACTAGT GACACTCGA GGATCGAG	Получение образца картирования для МПРА
29	Libr-rev	GTCTCGTGGGCTCGGAG ATGTGTATAAGAGACA	Получение образца картирования для МПРА

		GCCCTAGAAAGATAAT CATATTGT	
30	Libr-cDNA- for	GTCTCGTGGGCTCGGAG ATGTGTATAAGAGACA GGTCCTGCTGGAGTTCG TGAC	Получение образцов нормирования и экспрессии для МПРА
31	Libr_cDNA _A10_rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG AGTCGCCG CGCCAGGG TTTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
32	Libr_cDNA _A11_rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG T AAACATC CGCCAGGGTT TTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
33	Libr_cDNA _A12_rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG ACAATTCG CGCCAGGGT TTTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
34	Libr_cDNA _A13_rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG T ACTTGTC CGCCAGGGTT TTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
35	Libr_cDNA _A14_rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG GTACCGTT CGCCAGGGT TTTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
36	Libr-cDNA- A15-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG	Получение образцов нормирования и экспрессии для МПРА

		GCCACATACGCCAGGG TTTTCCCAGTCACAAGG	
37	Libr-cDNA-A16-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG CCTATGGTCGCCAGGGT TTTTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
38	Libr-cDNA-A17-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG AACGTCGCGGCCAGGG TTTTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
39	Libr-cDNA-A18-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG AGGCAGCAGGCCAGGG TTTTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
40	Libr-cDNA-A19-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG AGCTTTCTCGCCAGGGT TTTTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
41	Libr-cDNA-A20-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG GGTATGTCGCCAGGGT TTTTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
42	Libr-cDNA-A21-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG GAGGGACCGGCCAGGG TTTTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
43	Libr-cDNA-A22-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAGT	Получение образцов нормирования и экспрессии для МПРА

		AGCTCTACGCCAGGGTT TTCCCAGTCACAAGG	
44	Libr-cDNA-A23-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAGT AATTGCGCGCCAGGGTT TTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
45	Libr-cDNA-A24-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG GAAATGGGCGCCAGGG TTTTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
46	Libr-cDNA-A25-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAGT CGAGACTCGCCAGGGTT TTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
47	Libr-cDNA-A26-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG CAGAGAGGCGCCAGGG TTTTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
48	Libr-cDNA-A27-rev	TCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG ATTAGTCACGCCAGGGT TTCCCAGTCACAAGG	Получение образцов нормирования и экспрессии для МПРА
49	Libr-P5-for	AATGATACGGCGACCA CCGAGATCTACACTCGT CGGCAGCGTC	Присоединение адаптера P5 для секвенирования образцов МПРА на платформе Illumina
50	Libr-P7-rev	CAAGCAGAAGACGGCA TACGAGATGTCTCGTGG GCTCGG	Присоединение адаптера P7 для секвенирования образцов МПРА на платформе Illumina

51	Illumina-qPCR-1	AATGATACGGCGACCA CCGA	кПЦР образцов МПРА перед NGS
52	Illumina-qPCR-2	CAAGCAGAAGACGGCA TACGA	кПЦР образцов МПРА перед NGS
53	hPGK-XbaI-for	ATAATCTAGAACCTCG CAGACGGACAGC	Мутация сайта SpeI в промоторе hPGK
54	2936- TTTTCACT -F	CTAGTCATGCGTCAATT TTTTCACTGATTATCTTT AAC	Встройка мутации TTTTCACT в позицию +29..36 п.н. после СПА sNRP-1
55	2936- TTTTCACT -R	GTACGTAAAGATAATC AGTGAAAAAATTGACG CATGA	Встройка мутации TTTTCACT в позицию +29..36 п.н. после СПА sNRP-1
56	2936- GTCTCTC T-F	CTAGTCATGCGTCAATT GTCTCTCTGATTATCTTT AAC	Встройка мутации GTCTCTCT в позицию +29..36 п.н. после СПА sNRP- 1
57	2936- GTCTCTC T-R	GTACGTAAAGATAATC AGAGAGACAATTGACG CATGA	Встройка мутации GTCTCTCT в позицию +29..36 п.н. после СПА sNRP- 1
58	2936- GTGTA CT T-F	CTAGTCATGCGTCAATT GTGTA CTGATTATCTT TAAC	Встройка мутации GTGTA CTT в позицию +29..36 п.н. после СПА sNRP- 1
59	2936- GTGTA CT T-R	GTACGTAAAGATAATC AAGTACACAATTGACG CATGA	Встройка мутации GTGTA CTT в позицию +29..36 п.н. после СПА sNRP- 1

60	2936- AAGCAAA G-F	CTAGTCATGCGTCAATT AAGCAAAG GATTATCTT TAAC	Встройка мутации AAGCAAAG в позицию +29..36 п.н. после СПА sNRP- 1
61	2936- AAGCAAA G-R	GTACGTAAAGATAATC CTTTGCTT AATTGACGC ATGA	Встройка мутации AAGCAAAG в позицию +29..36 п.н. после СПА sNRP- 1
62	2936- GCACCCT T-F	CTAGTCATGCGTCAATT GCACCCTT GATTATCTT TAAC	Встройка мутации GCACCCTT в позицию +29..36 п.н. после СПА sNRP- 1
63	2936- GCACCCT T-R	GTACGTAAAGATAATC AAGGGTGC AATTGACG CATGA	Встройка мутации GCACCCTT в позицию +29..36 п.н. после СПА sNRP- 1
64	2936- TCAGATA C-F	CTAGTCATGCGTCAATT TCAGATAC GATTATCTT TAAC	Встройка мутации TCAGATAC в позицию +29..36 п.н. после СПА sNRP- 1
65	2936- TCAGATA C-R	GTACGTAAAGATAATC G TATCTGA AATTGACGC ATGA	Встройка мутации TCAGATAC в позицию +29..36 п.н. после СПА sNRP- 1
66	2936- ACACCCA T-F	CTAGTCATGCGTCAATT ACACCCAT GATTATCTT TAAC	Встройка мутации ACACCCAT в позицию +29..36 п.н. после СПА sNRP- 1

67	2936- ACACCCA T-R	GTACGTAAAGATAATC TAGGGTGT AATTGACGC ATGA	Встройка мутации ACACCCAT в позицию +29..36 п.н. после СПА sNRP- 1
68	2936- GCCGCAG A-F	CTAGTCATGCGTCAATT GCCGCAGA GATTATCTT TAAC	Встройка мутации GCCGCAGA в позицию +29..36 п.н. после СПА sNRP- 1
69	2936- GCCGCAG A-R	GTACGTAAAGATAATC TCTGCGGC AATTGACGC ATGA	Встройка мутации GCCGCAGA в позицию +29..36 п.н. после СПА sNRP- 1
70	2936- GACTGCA T-F	CTAGTCATGCGTCAATT GACTGCAT GATTATCTT TAAC	Встройка мутации GACTGCAT в позицию +29..36 п.н. после СПА sNRP- 1
71	2936- GACTGCA T-R	GTACGTAAAGATAATC ATGCAGTC AATTGACGC ATGA	Встройка мутации GACTGCAT в позицию +29..36 п.н. после СПА sNRP- 1
72	WT_30_T_t o_G_F	CTAGTCATGCGTCAATT T GACGCATGATTATCTT TAAC	Замена Т на Г в позиции +30 п.н. после СПА sNRP-1
73	WT_30_T_t o_G_R	GTACGTAAAGATAATC ATGCGT C AAATTGACGC ATGA	Замена Т на Г в позиции +30 п.н. после СПА sNRP-1
74	WT_33_G_t o_T_F	CTAGTCATGCGTCAATT TTACT T CATGATTATCTT TAAC	Замена G на T в позиции +33 п.н. после СПА sNRP-1

75	WT_33_G_to_T_R	GTACGTAAAGATAATC ATGAGTAAAATTGACG CATGA	Замена G на T в позиции +33 п.н. после СПА sNRP-1
76	SV40-EcoRI-fwd	AAAGAATTCCAACCTTGT TTATTGCAGCTTATAAT GG	Замена района -97..+92 п.н. после СПА sNRP-1 на терминатор SV40 early
77	SV40-MluI-rev	AAAACGCGTCAGACAT GATAAGATACATTGATG AGTTTG	Замена района -97..+92 п.н. после СПА sNRP-1 на на терминатор SV40 early
78	SV40_31_T_to_A-F	TAGTTGTGGTTTGACCA AACTCATCAATGTAT	Замена T на A в позиции +31 п.н. после СПА терминатора SV40 early
79	SV40_31_T_to_A-R	GATGAGTTTGGTCAAAC CACAACTAGAATG	Замена T на A в позиции +31 п.н. после СПА терминатора SV40 early
80	SV40_31_ins_T-F	GTTGTGGTTTGTCTCAA ACTCATCAATGTATC	Инсерция T в позиции +33 п.н. после СПА терминатора SV40 early
81	SV40_33_ins_T-R	TGATGAGTTTGAGACAA ACCACAAC TAGAA	Инсерция T в позиции +33 п.н. после СПА терминатора SV40 early
82	bGH-EcoRI	AAAGAATTCCTGTGCCT TCTAGTTGCCA	Замена района -97..+92 п.н. после СПА sNRP-1 на терминатор гена bGH
83	bGH-MluI	AATACGCGTTCCCCAGC ATGCCTGCTATTC	Замена района -97..+92 п.н. после СПА sNRP-1 на терминатор гена bGH

84	bGH_21_ins _A-F	CATCGCATTTGTCTAGAG TAGGTGTCATTCTAT	Инсерция А в позиции +21 п.н. после СПА терминатора гена bGH
85	bGH_21_ins _A-R	GACACCTACTCTAGACA ATGCGATGCAATT	Инсерция А в позиции +21 п.н. после СПА терминатора гена bGH
86	bGH_25_A _to_T-F	GCATTGTCTGAGTTGGT GTCATTCTATTCTGG	Замена А на Т в позиции +25 п.н. после СПА терминатора гена bGH
87	bGH_25_A _to_T-R	GAATGACACCAACTCA GACAATGCGATG	Замена А на Т в позиции +25 п.н. после СПА терминатора гена bGH
88	3'-eGFP-F	CATGGACGAGCTGTAC AAGTAAG	Замена одиночных нуклеотидов в терминаторах транскрипции
89	PacI-R	TAAAGGGAACAAAAGC TGGATTAAT	Замена одиночных нуклеотидов в терминаторах транскрипции, ОТ
90	Nluc-SmaI- for	ATAACCCGGGACCATG GTCTTCACACTCG	Амплификация последовательности люциферазы (NanoLuc)
91	Nluc-EcoRI- rev	ATATGAATTCTTACGCC AGAATGCGTTCG	Амплификация последовательности люциферазы (NanoLuc)
92	олиго(dT) ₂₀	NVTTTTTTTTTTTTTTTTT TTTT	Синтез кДНК
93	hPGK1- qPCR-F1	CCCTTCCTGGCCATCCT G	Оценка количества зрелой мРНК hPGK1 методом кПЦР

94	hPGK1-qPCR-R1	TCATCTCATTGACTTTG TCCAGC	Оценка количества зрелой мРНК hPGK1 методом кПЦР
95	qGFP-DpnI-F1	CTTCAAGATCCGCCACA ACATC	Оценка количества зрелой мРНК eGFP методом кПЦР
96	qGFP-DpnI-R1	GGACCATGTGATCGCGC TTCTC	Оценка количества зрелой мРНК eGFP методом кПЦР
97	mCher-qPCR-F2	TCACCTTCAGCTTGGCG GTCT	Оценка количества зрелой мРНК mCherry методом кПЦР
98	mCher-qPCR-R2	ATCAAGGAGTTCATGCG CTTCAAG	Оценка количества зрелой мРНК mCherry методом кПЦР
99	EGFP-int3	TCAAGGTGAACTTCAAG ATCCG	Секвенирование, ПЦР с колоний
100	pBS-R	TGACTTGAGCGTCGATT TTTG	Секвенирование, ПЦР с колоний
101	BC1	ACTAGGCAAGGCACCA G	Праймер к ШК контрольной плазмиды №1
102	BC2	CGAGGGATATAGAGCG AGTTAA	Праймер к ШК контрольной плазмиды №2
103	ROI1	TACGTTAAAGATAATCA TGCATAAGTCG	Праймер к мутации контрольной плазмиды №1
104	ROI2	TACGTTAAAGATAATCA TGCTATATGGC	Праймер к мутации контрольной плазмиды №2