УДК 577.21

Классификация и количественная оценка событий непродуктивного сплайсинга

Л. Г. Завилейский*, Е. А. Чернявская, М. А. Власенок, Д. Д. Первушин
Центр молекулярной и клеточной биологии, Сколковский институт науки и технологий, Москва, 121205 Россия
*E-mail: l.zavileisky@skoltech.ru
Поступила в редакцию 26.11.2024
Принята к печати 31.03.2025
DOI: 10.32607/actanaturae.27572

РЕФЕРАТ **В эукариотических клетках мРНК, содержащие преждевременные стоп-кодоны, разрушаются системой нонсенс-опосредованного распада (NMD). В результате взаимодействия между системами NMD и альтернативного сплайсинга генерируются NMD-чувствительные транскрипты (NMD targets, NMDT), которые играют важную роль в регуляции экспрессии генов по механизму непродуктивного сплайсинга. Для понимания этого механизма необходимо правильно идентифицировать события альтернативного сплайсинга, приводящие к появлению NMDT. В данной работе для нахождения событий альтернативного сплайсинга, их классификации и количественной оценки разработан вычислительный конвейер NMDj, который в отличие от существующих методов не опирается на сравнение NMDT с наиболее похожим на него кодирующим транскриптом, а использует набор характеристических интронов, отличающих NMDT от кодирующих транскриптов. Тестирование на смоделированных данных секвенирования РНК показало, что NMDj способен количественно определять события альтернативного сплайсинга, приводящие к проявлению NMDT, с большей точностью, чем другие существующие для этой цели методы. NMDj представляет собой универсальный метод, подходящий для классификации сколь угодно сложных событий альтернативного сплайсинга, приводящих к появлению NMDT. Вычислительный конвейер NMDj доступен через репозиторий https://github.com/zavilev/NMDj/.**

КЛЮЧЕВЫЕ СЛОВА **непродуктивный сплайсинг, нонсенс-опосредованный распад, NMD, сплайсинг.**

СПИСОК СОКРАЩЕНИЙ **NMD – нонсенс-опосредованный распад (Nonsense Mediated Decay); NMDT – транскрипт-мишень NMD (NMD target); АС – альтернативный сплайсинг; ПСК – преждевременный стоп-кодон; НТО – нетранслируемая область; нт – нуклеотид.**

ВВЕДЕНИЕ

Благодаря альтернативному сплайсингу (АС) гены эукариот могут экспрессировать большое число изоформ транскриптов. По грубым оценкам, гены человека, кодирующие белки, на детектируемом уровне экспрессии производят до ~150000 различных транскриптов, в среднем по 7.4 изоформы на ген [1]. Однако только половина из них кодирует полноразмерные белки, а остальные могут содержать преждевременные стоп-кодоны (ПСК) [1, 2]. У эукариот такие транскрипты избирательно уничтожаются системой, называемой нонсенс-опосредованным распадом (NMD) [3].

NMD не только предотвращает трансляцию усеченных белков, возникающих в результате нонсенс-мутаций и ошибок сплайсинга, но также участвует в различных биологических процессах, включая регуляцию экспрессии генов [4]. Большинство РНК-связывающих белков контролируют уровень собственной экспрессии посредством петли отрицательной обратной связи, в которой белковый продукт гена связывается с кодирующей его мРНК и индуцирует в ней АС, который приводит к появлению ПСК [5, 6]. Считается, что АС в большей степени воздействует на транскрипционный ландшафт эукариот путем генерации NMD-изоформ для ограничения уровней экспрессии генов, чем путем расширения разнообразия протеома [2].

Локальные, то есть сосредоточенные на небольшом участке пре-мРНК, изменения сплайсинга являются одним из основных источников транскриптов, служащих мишенями NMD (NMD targets, NMDT). Среди основных типов локальных событий АС можно выделить АС так называемых ядовитых и необходимых экзонов, которые приводят к появлению NMDT при включении и пропуске экзона соответственно, а также использование альтернативных 5'- и 3'-сайтов сплайсинга и удержание ин-

тронов [7]. Некоторые из них могут быть вовлечены в регуляцию одного и того же биологического процесса (например, удержания интронов) или регулироваться одним и тем же фактором сплайсинга [8, 9]. Однако многообразие всех событий АС не исчерпывается перечисленными основными типами [6]. Задача характеризации сложных событий АС, приводящих к появлению NMDT, возникает во многих исследованиях, связанных с изучением регуляции генной экспрессии [10–12].

На сегодняшний день решить эту задачу можно только с помощью программы NMD Classifier [13]. Используемый в этой программе подход основан на предположении о минимальной эволюции/регуляции, согласно которому NMDT возникают в результате эволюционных или регуляторных событий, которые наименьшим возможным образом изменяют рамку считывания кодирующего транскрипта. То есть, NMD Classifier для каждого NMDT находит наиболее похожий (с точки зрения общей нуклеотидной последовательности) кодирующий транскрипт и считает искомым событием АС различия между найденным наилучшим транскриптом-партнером и NMDT, вызывающие сдвиг рамки считывания. Однако вероятность того, что NMDT был получен из кодирующего транскрипта в результате АС, зависит не только от сходства их экзон-интронной архитектуры, но и от уровня экспрессии. Кодирующий транскрипт с самым высоким уровнем экспрессии с большей вероятностью будет источником NMDT [14]. Более того, NMDT может быть получен из разных транскриптов со сравнимыми уровнями экспрессии, что ставит под сомнение обоснованность подхода, состоящего в выборе единственного лучшего транскрипта-партнера.

Возвращаясь к этой задаче, мы разработали NMDj – инструмент для систематического поиска, классификации и количественной оценки событий АС, приводящих к появлению NMDT. NMDj учитывает все аннотированные транскрипты и сообщает обо всех альтернативных интронах, отличающих NMDT от кодирующих транскриптов. NMDj обеспечивает более подробную классификацию событий АС, приводящих к появлению NMDT, чем NMD Classifier. Сопряжение между NMD и АС является важнейшим посттранскрипционным механизмом регуляции экспрессии генов [15]. Поэтому разработка метода поиска, классификации и количественной оценки событий АС, приводящих к появлению NMDT, с учетом всего многообразия изоформ транскриптов, является актуальной задачей. На ее решение и направлен метод NMDj. Он получает на вход множество транскриптов в виде аннотационной базы данных или моделей транскриптов, построенных по данным секвенирования РНК, а на выходе характеризует события АС, приводящие к появлению NMDT, и производит их количественную оценку.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Геномы и их аннотация

Аннотации геномов человека (GRCh38, версия 108), мыши (mm10, версия 113), данио-рерио (danRer11, версия 113) и дрозофилы (dm6, версия 113) были загружены из Ensembl в формате GTF [16]. При поиске событий АС, приводящих к появлению NMDT, рассматривали только транскрипты генов, кодирующих белки, по крайней мере с одним аннотированным NMDT. Транскрипты без аннотированных старт- или стоп-кодонов не рассматривали. Гены без NMDT или кодирующих транскриптов также не рассматривали.

NMD Classifier

Исходный код NMD Classifier был загружен с сайта [13]. Для количественной оценки локальных изменений сплайсинга выходные данные NMD Classifier были преобразованы в список альтернативных интронов, соответствующих четырем основным типам событий АС: альтернативные экзоны, альтернативные 5’- и 3’-сайты сплайсинга и удержание интронов (NMD_in, NMD_ex, A5SS, A3SS, NMD_IR, nNMD_IR).

Вычислительный конвейер NMDj

Входными данными для NMDj является аннотация транскриптома в формате GFF/GTF [17]. Рассматриваются следующие четыре ее элемента («transcript», «exon», «start_codon» и «stop_codon») и три атрибута («gene_id», «transcript_id», «transcript_biotype»). В дополнение к основному файлу GFF/GTF NMDj также может принимать файл с аннотацией транскриптов, содержащий элементы «transcript» и «exon» и атрибут «transcript_id». В этом случае каждый транскрипт из дополнительного файла приписывается гену из основного файла на основе максимального количества общих интронов и перекрытия последовательности не менее 50%. Для транскриптов, которые были приписаны генам, выбирается самая длинная рамка считывания из числа тех, которые начинаются с аннотированных для гена старт-кодонов, и к аннотации добавляются соответствующие позиции старт- и стоп-кодонов. Как и в Ensembl [18], транскрипт аннотируется как NMDT, если в нем на расстоянии не менее 50 нт в направлении 3’-конца от стоп-кодона присутствует интрон.

Затем для каждого NMDT рассматривается геномный интервал от последнего сайта сплайсинга,

общего для NMDT и любого кодирующего транскрипта с той же фазой (или старт-кодона, в отсутствие такового), до 3'-конца экзона с ПСК или ближайшего конца транскрипта, если NMDT имеет общий стоп-кодон с кодирующим транскриптом. Характеристические интроны определяются как все интроны, пересекающиеся с найденным геномным интервалом, за исключением тех, которые являются общими для NMDT и любого кодирующего транскрипта. Событием АС, приводящим к появлению NMDT, считается набор найденных характеристических интронов. События АС из пары NMDT объединяются в кластер, если данные NMDT имеют хотя бы один общий характеристический интрон.

Для того, чтобы классифицировать события АС, приводящие к появлению NMDT, NMDj по умолчанию использует транскрипты MANE-Select в качестве эталона, поскольку они, как правило, являются наиболее высоко экспрессируемыми [19]. Однако пользователь может задать и другое множество транскриптов сравнения. NMDj строит ориентированный ациклический граф, используя сайты сплайсинга NMDT и транскрипта сравнения в качестве узлов, а интроны и экзоны в качестве ребер, и ищет в нем «пузыри» (bubbles), определенные вершинно-независимыми путями, которые содержат характеристические интроны [20, 21]. NMDj выводит найденные пары вершинно-независимых путей в виде $X_1...X_n : Y_1..Y_m$, где X_i и Y_j – это символы «D» (донор) или «A» (акцептор). При этом $X_i \neq X_j$ и $Y_i \neq Y_j$, когда $j = i \pm 1$. Если множество транскриптов сравнения не было задано, то NMDj итеративно сравнивает NMDT с каждым кодирующим транскриптом.

Последним, необязательным, шагом является количественная оценка АС с использованием чтений с разрывами из экспериментов по секвенированию РНК (подаются на вход в виде таблицы). NMDj вычисляет значения метрики сплайсинга Ψ, которая оценивает уровень экспрессии NMDT относительно экспрессии всех транскриптов гена. Она рассчитывается по формуле:

$$\Psi = \frac{\sum_{i=1}^{A} a_i k_i}{\sum_{i=1}^{A} a_i k_i + \sum_{j=1}^{B} b_j r_j},$$

где A и B – число характеристических интронов, поддерживающих NMDT и кодирующие транскрипты, a_i и b_j – число чтений с разрывами, соответствующих характеристическим интронам, а k_i и r_j – веса, которые учитывают кратности, с которыми характеристические интроны встречаются в NMDT и кодирующих транскриптах соответственно. Веса k_i и r_j находятся независимо для NMDT и кодирующих транскриптов. Естественно потребовать, чтобы сумма весов характеристических интронов, входящих в каждый транскрипт, была равна единице. Это приводит нас к системе из n линейных уравнений с m неизвестными, где n – число транскриптов, а m – число характеристических интронов. По построению характеристических интронов такая система всегда совместна, однако она может иметь бесконечное число решений. В общем случае однозначный выбор k_i и r_j можно сделать, накладывая на эту систему ограничения регуляризации. Однако в NMDj мы используем следующий эвристический алгоритм, который позволяет определить значение Ψ в соответствии с данными ранее определениями для основных типов событий АС [6, 11].

Аннотированные в найденном интервале транскрипты представляются в виде графа, вершинами которого являются характеристические интроны, а ребрами – соединяющие их экзоны (или их группы). Затем в этом графе происходит поиск пар вершин, между которыми существуют только вершинно-независимые пути. Для каждого такого пути веса характеристических интронов в нем полагаются равными друг другу. Например, для ядовитого экзона таких путей будет два: один, соответствующий включению экзона (с двумя характеристическими интронами, каждый с весом 0.5), и другой, соответствующий пропуску экзона (с одним характеристическим интроном, вес которого равен 1). После того как коэффициентам вершин на путях между выбранной парой присвоены значения, эти вершины объединяются в одну и поиск продолжается в новом графе. На каждом шаге коэффициенты характеристических интронов, объединенных в вершину, домножаются на присвоенное ей значение, и процедура повторяется до тех пор, пока все вершины не будут объединены в одну. Данный алгоритм подходит для всех простых типов событий АС, а для сложных событий работает в предположении о вложенности вершинно-независимых путей.

Данные секвенирования РНК и их симуляция

Для того чтобы реалистично смоделировать данные секвенирования РНК, используя заведомо известные уровни экспрессии транскриптов, а значит и относительные уровни экспрессии NMDT, мы случайным образом выбрали по три образца из каждой из трех тканей (мышца, печень, мозжечок) в панели транскриптомных данных консорциума Genotype-Tissue Exression project (GTEx) [22]. Выбор тканей был обусловлен тем, что они наиболее значительно различаются по АС [23, 24]. Уровни экспрессии транскриптов в выбранных образцах были вычислены с помощью

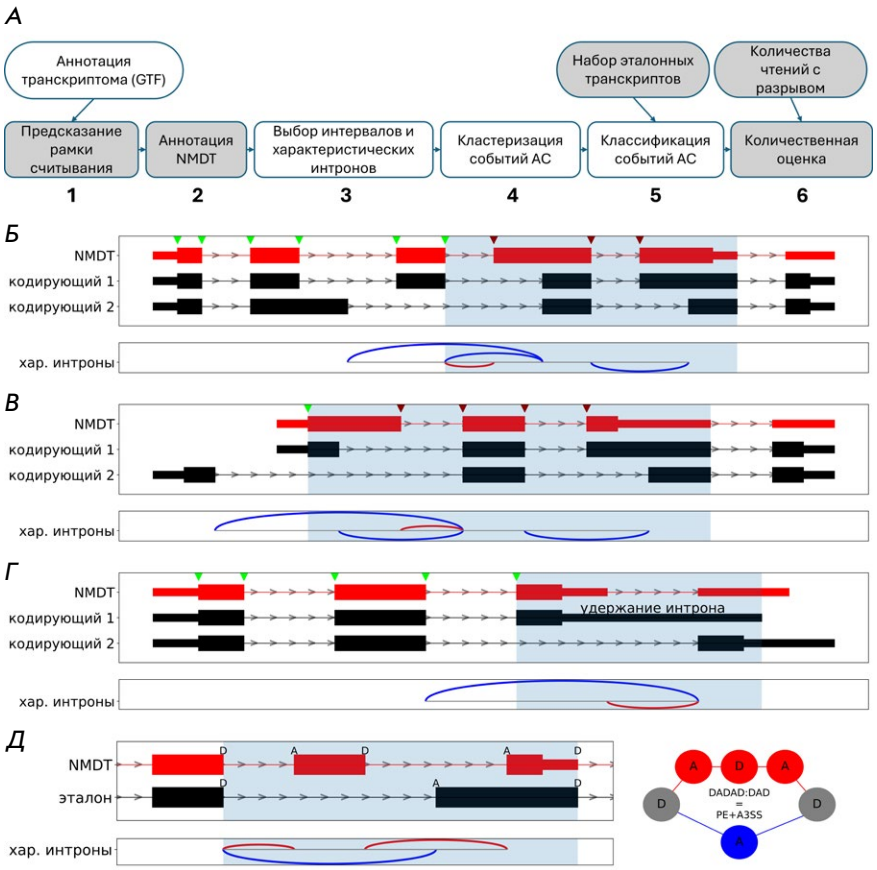

Рис. 1. Вычислительный конвейер NMDj. *А* – блок-схема конвейера. *Б*–*Г* – выбор границ интервала (обозначен голубой заливкой). 5′-Границей является либо последний сайт сплайсинга, общий для NMDT и любого кодирующего транскрипта с той же фазой (*Б*), либо стартовый кодон, если такого сайта сплайсинга нет (*В*). 3′-Граница представляет собой либо донорный сайт сплайсинга интрона, следующего за ПСК-содержащим экзоном (*Б*, *В*), либо конец самой короткой 3′-НТО после стоп-кодона NMDT (*Г*). *Д* – пример классификации на основе вершинно-независимых путей. NMDT и его эталонный кодирующий транскрипт (слева) соответствуют паре вершинно-независимых путей, состоящих из донорных и акцепторных сайтов сплайсинга (справа). NMDT и кодирующие транскрипты, а также соответствующие им характеристические интроны (дуги) обозначены красным и синим цветом соответственно. Сайты сплайсинга NMDT обозначены зелеными стрелками, если рамка NMDT соответствует кодирующей рамке, или красными в противном случае

программы rsem-calculate-expression с опцией estimate-rspd [25]. Для каждого гена рассчитывали доли экспрессии NMDT, наилучших транскриптов-партнеров, найденных NMD Classifier, и транскриптов MANE-Select от общей экспрессии гена. Выборку повторяли 5 раз, а результаты усредняли.

Симуляцию экспериментов по секвенированию РНК осуществляли с помощью программы rsem-simulate-reads с использованием уровней экспрессии транскриптов, полученных ранее. Для каждого образца было симулировано 50 млн парноконцевых чтений. Симулированные чтения были картированы на геном человека GRCh38 с использованием программы STAR Aligner 2.7.3a [26]. Чтения с разрывами были подсчитаны программой IPSA с настройками по умолчанию [27]. Уровни экспрессии транскриптов в симулированных данных определяли количественно с помощью RSEM (как указано выше) [25], Salmon 1.10.3 с опциями --seqBias --gcBias --posBias [28] и StringTie 2.2.3 с опцией -e [29]. Чтобы преобразовать результаты количественного определения уровней экспрессии транскриптов в значения Ψ, уровни экспрессии NMDT (в TPM, transcripts per million) были разделены на сумму уровней экспрессии транскриптов, пересекающих найденные NMDj геномные интервалы.

Данные секвенирования РНК при инактивации NMD

Результаты экспериментов по инактивации компонентов NMD (двойной нокдаун SMG6 и SMG7) с последующим секвенированием РНК получены из Gene Expression Omnibus под номером доступа GSE86148 в формате FASTQ и выровнены на сборку генома человека GRCh38 (hg38) с использованием STAR Aligner v2.7.8a в парноконцевом режиме. Чтения с разрывами были подсчитаны программой IPSA с настройками по умолчанию [27].

РЕЗУЛЬТАТЫ

Вычислительный конвейер NMDj

NMDj состоит из трех основных и трех вспомогательных этапов (*рис. 1А*). Получая на вход аннотацию транскриптов, программа выполняет поиск рамки считывания и предсказывает NMDT, если они не были аннотированы. Аннотация NMDT производится на основании так называемого правила 50 нт, которое постулирует, что транскрипт распознается системой NMD как мишень, если он содержит интрон на расстоянии не менее 50–55 нт в направлении 3’-конца от стоп-кодона [30]. Оно основано на предположении о том, что связанные вблизи экзон-эк-

Таблица 1. Список основных типов событий NMDj и их синонимов в классификации, предоставленной NMD Classifier

Тип	Событие	Описание	Синоним
DADA:DA	PE	Ядовитый кассетный экзон вызывает NMD при включении	NMD_in
D(AD)nA:DA	PEn	Одновременное включение *n* последовательных кассетных экзонов вызывает NMD	multi_NMD_in
DA:DADA	EE	Необходимый кассетный экзон вызывает NMD при пропуске	NMD_ex
DA:D(AD)nA	EEn	Одновременный пропуск *n* последовательных кассетных экзонов вызывает NMD	multi_NMD_ex
ADA:ADA	A5SS	Альтернативный 5'-сайт сплайсинга	A5SS
DAD:DAD	A3SS	Альтернативный 3'-сайт сплайсинга	A3SS
ADAD:ADAD	A5SS+A3SS	Альтернативные 5'- и 3'-сайты сплайсинга одного и того же интрона	A5SS, A3SS
AD:ADAD	IR	Удержание интрона вызывает NMD	nNMD_IR
ADAD:AD	ID	Вырезание интрона вызывает NMD	NMD_IR
DADA:DADA	MXE	Пара взаимоисключающих соседних экзонов	-
AD(AD)nA:ADA	A5SS+PEn	Альтернативный 5'-сайт сплайсинга и *n* последовательных ядовитых экзонов	-
ADA:AD(AD)nA	A5SS+EEn	Альтернативный 5'-сайт сплайсинга и *n* последовательных необходимых экзонов	-
D(AD)nAD:DAD	PEn+A3SS	*n* последовательных ядовитых экзонов и альтернативный 3'-сайт сплайсинга	-
DAD:D(AD)nAD	EEn+A3SS	*n* последовательных необходимых экзонов и альтернативный 3'-сайт сплайсинга	-
ADAD:AD(AD)nAD	A5SS+EE+A3SS	Альтернативный 5'-сайт сплайсинга, *n* последовательных необходимых экзонов и альтернативный 3'-сайт сплайсинга	-

зонных соединений белковые комплексы смещаются рибосомой во время первого раунда трансляции, а те, которые остаются связанными за пределами рамки считывания, служат сигналом о появлении ПСК [30]. В NMDj мы использовали порог в 50 нт, поскольку это значение принято для автоматической аннотации NMDT в Ensembl [16]. Однако количество предсказанных NMDT меняется незначительно при увеличении порога до 55 нт (*рис. S1*).

После того как для каждого гена предсказаны NMDT, программа находит события AC, приводящие к их появлению. Существуют различные формализации для описания событий AC, включая бинарную классификацию (например, ядовитые экзоны [31]), классификацию связных компонент в графе сплайсинга [32] и нахождение локальных вариантов сплайсинга [33]. В этой работе мы определяем событие AC как набор характеристических интронов, охватывающих определяемый далее геномный интервал. Для каждого NMDT 5'-граница интервала определяется как последний сайт сплайсинга, в котором фаза NMDT совпадает с фазой любого кодирующего транскрипта (*рис. 1Б*). Если такого сайта не существует, то 5'-границей считается старт-кодон NMDT, если он является общим хотя бы с одним кодирующим транскриптом (*рис. 1В*). 3'-Границей интервала считается 3'-конец экзона, содержащего ПСК, или, если NMDT имеет общий стоп-кодон с кодирующим транскриптом (и в этом случае это не настоящий ПСК), то концом интервала считается ближайший конец транскрипта (*рис. 1Г*).

Затем NMDj выбирает характеристические интроны, которые отличают NMDT от кодирующих транскриптов. Все интроны, которые примыкают к интервалу или пересекают его, за исключением общих для NMDT, и хотя бы одного кодирующего транскрипта, считаются характеристическими. В результате каждый NMDT характеризуется набором интронов, которые присутствуют либо в нем, либо в кодирующих транскриптах (*рис. 1Б,Г*, красные и синие дуги). Поскольку несколько NMDT часто имеют одинаковые или очень похожие наборы характеристических интронов, найденные события AC объединяются в кластеры, чтобы уменьшить избыточность.

NMDj классифицирует события сплайсинга на основные типы, такие как ядовитые (poison exon, PE) и необходимые (essential exon, EE) экзоны, альтернативные сайты сплайсинга (A5SS, A3SS) и другие (*табл. 1*). Классификация событий AC основана на концепции вершинно-независимых путей в применении к графу сплайсинга [20, 34]. В ориентированном ациклическом графе, вершинами которого являются донорные (D) и акцепторные (A) сайты

Таблица 2. Приводящие к появлению NMDT события AC в транскриптах человека и модельных организмов

Организм	#Tr	#NMDT	NMDT, %	Доля событий AC, %					
				PE	EE	A5SS	A3SS	IR	Другие
Человек	79940	16741	21	18	11	6	8	2	55
Мышь	49951	5339	11	18	18	11	14	4	36
Данио-рерио	35040	854	2	11	10	11	12	23	32
Дрозофила	30688	1325	4	18	4	12	9	16	41

Примечание. #Tr – общее число транскриптов; #NMDT – число NMDT; NMDT – доля NMTD в %. Доли (в %) ядовитых (PE) и необходимых (EE) экзонов, доли альтернативных 5′-(A5SS) и 3′- сайтов сплайсинга (A3SS), доля удержанных интронов (IR) и доля остальных событий (Другие).

сплайсинга, а ребрами – экзоны и интроны, вершинно-независимые пути можно определить как пару путей, которые не имеют общих вершин, кроме первой и последней (*рис. 1Д*). Каждая такая пара представляется в символической форме, соответствующей последовательности вершин, т.е. ядовитый экзон соответствует DADA:DA, альтернативный 5'-сайт сплайсинга (A5SS) – ADA:ADA, а множественные ядовитые экзоны (PEn) – D(AD)nA:DA, где n – количество экзонов. На последнем этапе NMDj оценивает долю экспрессии NMDT в экспрессии гена в виде метрики Ψ, вычисленной на основании числа чтений с разрывами из экспериментов секвенирования РНК (см. «Экспериментальную часть»).

Применение NMDj к транскриптам человека и модельных организмов

Применение NMDj к аннотированным транскриптам человека, мыши, данио-рерио и дрозофилы показало, что доля NMDT, подчиняющихся правилу 50 нт, у человека и мыши существенно выше, чем у данио-рерио и дрозофилы, что несомненно является результатом различий в качестве и полноте аннотаций транскриптомов (*табл. 2*). Однако частоты событий AC, приводящих к появлению NMDT, значительно различаются между организмами. Если у человека и мыши появление NMDT происходит в большей степени за счет использования ядовитых и необходимых экзонов, чем за счет удержания интронов, то у дрозофилы и данио-рерио наблюдается обратная тенденция. По существующим оценкам доля удержания интронов среди основных типов AC одинаково низка как у млекопитающих, так и у других позвоночных и беспозвоночных [35]. Таким образом, наблюдаемое различие между частотами приводящих к появлению NMDT событий AC не может объясняться ни разной представленностью их типов, ни разной полнотой аннотации транскриптома, а скорее указывает на особенности функционирования системы NMD в разных таксономических группах.

Преимущества NMDj для нахождения событий AC, приводящих к NMD

Существующий подход к анализу событий AC, приводящих к появлению NMDT, реализованный в методе NMD Classifier, основан на нахождении наилучшего транскрипта-партнера. Его главной проблемой является то, что при выборе наилучшего партнера не учитываются другие транскрипты и уровни их экспрессии. Кодирующий транскрипт вряд ли может быть основным источником NMDT, если уровень его экспрессии низок. Чтобы проиллюстрировать важность этой проблемы, мы применили NMD Classifier к аннотации транскриптома Ensembl [16] и сравнили множество найденных транскриптов-партнеров с транскриптами из аннотации MANE-Select в качестве множества наиболее экспрессируемых транскриптов в каждом гене человека [19].

Транскрипты MANE-Select определены как наилучшие транскрипты-партнеры только для 25% NMDT, хотя они имели значительно более высокий уровень экспрессии, что подтверждается случайной выборкой экспериментов по секвенированию РНК из GTEx (*рис. 2А*). Более того, когда наилучший транскрипт-партнер не совпадал с MANE-Select, его вклад в общий уровень экспрессии генов был сопоставим с вкладом NMDT. Это указывает на то, что кодирующий транскрипт с наибольшей длиной общей с NMDT последовательности может не быть транскриптом, из которого NMDT получается в результате AC. Кроме того, транскрипты MANE-Select не всегда являются наиболее экспрессируемыми. Ткани могут различаться по наиболее экспрессируемым транскриптам (*рис. 2Б*) или экспрессировать несколько транскриптов на сопоставимых уровнях. Поэтому NMDj рассматривает все аннотированные транскрипты для того, чтобы избежать проблемы выбора одного наилучшего транскрипта-партнера, и кластеризует NMDT по характеристическим интронам, чтобы получить краткий и неприводимый набор событий AC.

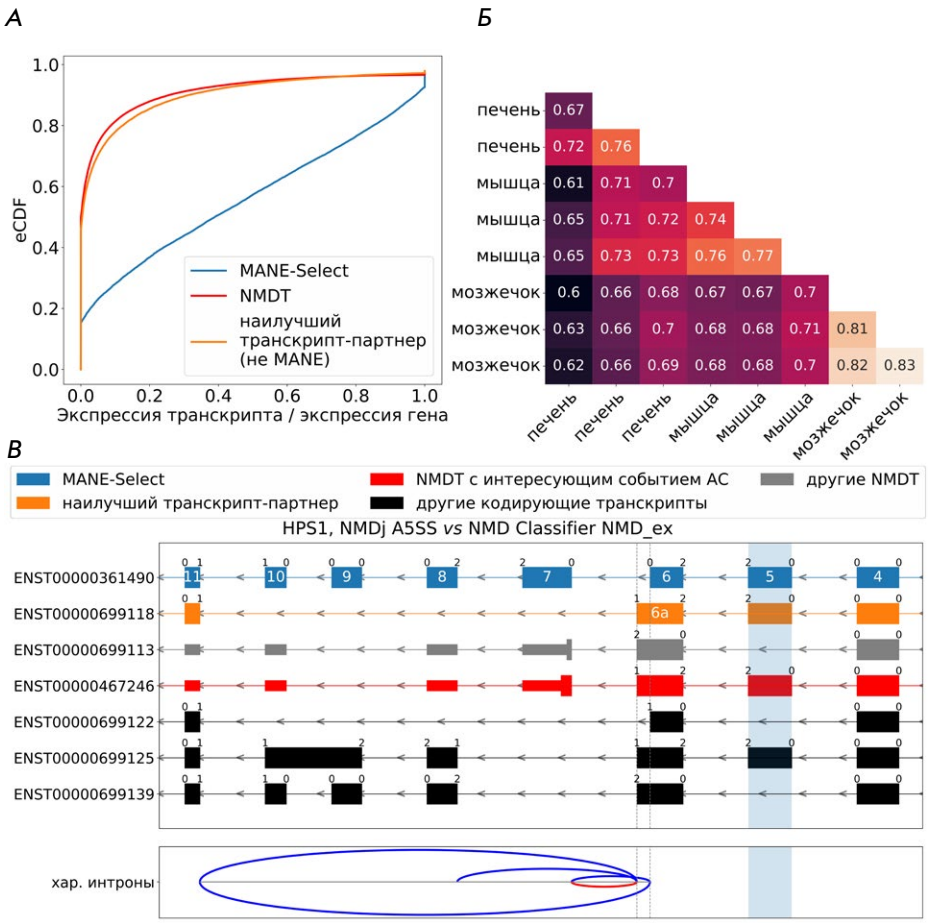

Рис. 2. NMDj и наилучшие транскрипты-партнеры NMD Classifier. *А* — относительные доли транскриптов (eCDF – кумулятивная функция распределения) изоформ MANE-Select, изоформ наилучшего транскрипта-партнера и изоформ NMDT, оцененные по выборке экспериментов секвенирования РНК из тканей человека. *Б* – доля генов, наиболее экспрессируемые транскрипты которых совпадают между парами образцов тканей. *В* – пример локального события AC, приводящего к появлению NMDT, в гене *HPS1*. Характеристические интроны для NMDT и кодирующих транскриптов показаны красными и синими дугами соответственно. Фаза рамки считывания указана над границами экзонов. Цвета транскриптов: MANE-Select (синий), NMDT (красный), наилучший транскрипт-партнер, найденный NMD Classifier (оранжевый), другие транскрипты (серый – NMDT, черный – кодирующие). Необходимый экзон, обнаруженный NMD Classifier, выделен вертикальной голубой заливкой, однако в действительности NMDT образуется из-за смещения сайта сплайсинга в MANE-Select-изоформе

NMDj особенно полезен в генах со сложной архитектурой сплайсинга. Ярким примером является ген *HPS1*, который содержит группу экзонов, длина которых не кратна трем (*рис. 2В*). Пропуск каждого отдельного экзона приводит к появлению NMDT, если он не компенсируется событием AC в направлении 3'-конца, которое восстанавливает рамку считывания. Одновременное включение экзонов 6а и 7 приводит к появлению NMDT. NMD Classifier выбирает транскрипт с экзоном 5 в качестве наилучшего транскрипта-партнера. Этот экзон пропускается в NMDT, что действительно вызывает сдвиг рамки считывания. Однако он также пропускается в кодирующем транскрипте, в котором его сдвиг рамки компенсируется использованием экзона 6 вместо экзона 6а и пропуском экзонов 7–10. NMDj правильно идентифицирует последний сайт сплайсинга, в котором рамка считывания NMDT совпадает с рамкой считывания кодирующего транскрипта, как 3'-границу экзона 6а, что позволяет обнаружить единственное истинное событие AC, приводящее к появлению NMD, а именно вырезание интрона между экзонами 6а и 7. Он также идентифицирует все альтернативные интроны, вырезание которых позволяет избежать сдвига рамки. Интересно, что другой NMDT с включенным экзоном 5 имеет тот же характеристический интрон и поэтому кластеризуется с предыдущим.

NMDj дает более подробную классификацию событий AC

Мы сравнили результаты классификации событий AC, полученные с помощью NMDj и NMD Classifier, в применении к одной и той же аннотации транскриптома человека (*рис. 3А*). NMDj настроен на использование транскриптов MANE-Select в качестве эталона. Из 15914 NMDT программы NMD Classifier и NMDj смогли классифицировать события AC для 15446 и 15265 NMDT соответственно. Однако события AC были отнесены к одному и тому же типу (*табл. 1*) только в 60% случаев.

В то время как NMD Classifier подразделяет события AC на небольшое число наиболее распространенных типов, NMDj может описывать более сложные события сплайсинга, используя вершинно-независимые пути. Например, в гене *POR* NMDT отличается от кодирующих изоформ альтернативными 5'- и 3'-сайтами сплайсинга и кассетным экзоном (*рис. 3Б*). Такие события обычно пропускают-

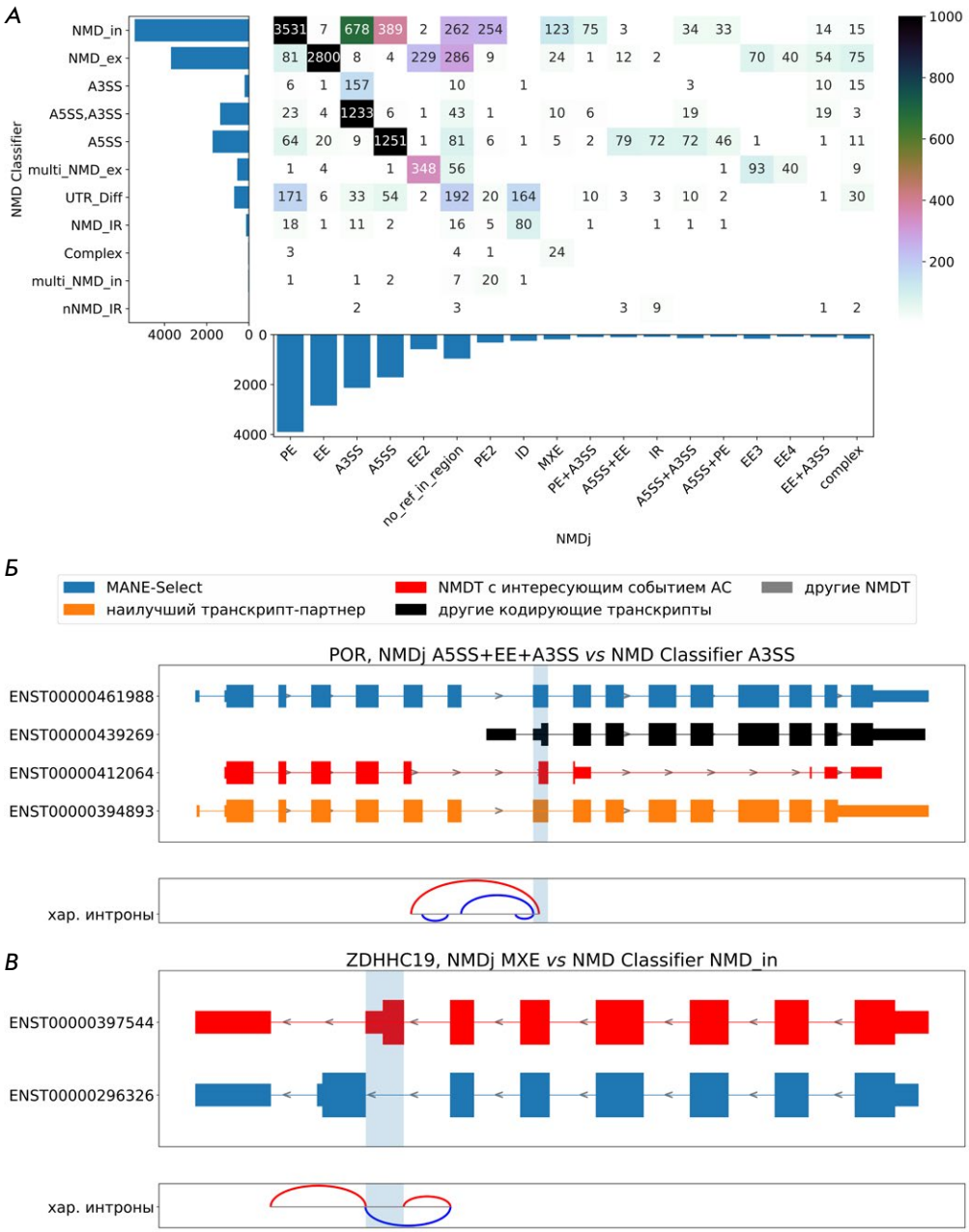

Рис. 3. Категоризация событий AC. *А* – сравнение классификаций NMDj и NMD Classifier. Каждая ячейка представляет количество NMDT, классифицированных на соответствующие типы с помощью NMDj (строки) и NMD Classifier (столбцы). *Б*, *В* – примеры редких событий AC, приводящих к образованию NMDT. Остальные обозначения как на *рис. 2*

ся многими стандартными методами исследования сплайсинга [31, 32]. Наличие типов AC, которые NMD Classifier не может правильно классифицировать, является причиной большой доли несоответствий между двумя классификациями. Например, ряд событий, определенных NMD Classifier как ядовитые экзоны (NMD_in), классифицируются NMDj как PE+A3SS и MXE (*рис. 3А,В*). Еще одно преимущество NMDj – способность классифицировать события AC в 3'-нетранслируемых областях (3'-НТО). Большинство событий, индуцирующих NMD в 3'-НТО, ожидаемо представлено удержанием интронов. Более того, многие события 3'-НТО не пересекаются с изоформой MANE-Select (*рис. 3А, S2*).

Относительно небольшое количество других несоответствий можно объяснить тем фактом, что NMDj и NMD Classifier используют разные эталонные транскрипты для классификации событий AC. Лишь 61% событий AC были классифицированы одинаково, когда NMDj был настроен на использование наилучших транскриптов-партнеров в качестве эталона. Значительная часть различий обусловлена неправильной классификацией NMD Classifier. Например, большинство событий, отнесенных NMD

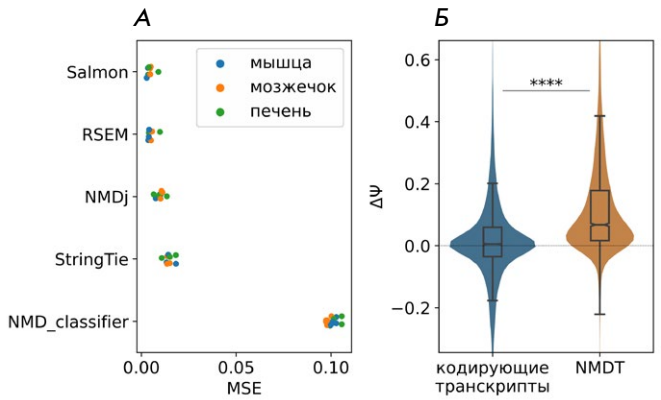

Рис. 4. Сравнение предсказаний NMDj и NMD Classifier. *А* – среднеквадратическая ошибка (MSE) значений Ψ, оцененных различными методами по симулированным данным секвенирования РНК относительно истинного значения. *Б* – статистически значимые изменения событий AC, приводящих к появлению NMDT и кодирующих транскриптов (кассетные экзоны, альтернативные сайты сплайсинга и удержание интронов) при инактивации NMD циклогексимидом. **** обозначает статистически значимые различия на 0.01% уровне значимости (тест Манна–Уитни)

Classifier к типу «A3SS, A5SS», классифицируются NMDj как A3SS (*рис. 3А*). Между тем, размер класса «A5SS,A3SS» NMD Classifier намного больше, чем размер класса «A3SS». Это противоречит здравому смыслу, поскольку выбор между парой альтернативных 5'-сайтов сплайсинга, по-видимому, не зависит от выбора между парой альтернативных 3'-сайтов сплайсинга, расположенных в направлении 3'-конца и разделенных длинным интроном [36]. Визуальная проверка случайно выбранных примеров несоответствия классификации подтверждает правильность классификации NMDj (*рис. S3*).

Валидация NMDj на симулированных и реальных данных

Приводящие к появлению NMDT события AC можно использовать для количественной оценки относительных уровней экспрессии NMDT с использованием данных секвенирования РНК. Чтобы оценить точность NMDj в количественной оценке AC, мы симулировали данные секвенирования РНК, используя средние уровни экспрессии транскриптов в тканях человека. Оценочные значения Ψ, которые были рассчитаны на основе чтений с разрывами, соответствующих характеристическим интронам, сравнивали с фактическими значениями Ψ, определяемыми как доля уровня экспрессии NMDT в суммарном уровне экспрессии всех транскриптов гена. В качестве меры расстояния использовали среднеквадратичную ошибку (MSE) по всем значениям Ψ для всех изоформ NMDT. Оказалось, что по точности NMDj сравним с современными методами количественной оценки на уровне транскриптов, в то время как значения MSE для NMD Classifier были значительно выше (*рис. 4А*). Поскольку способ вычисления метрики Ψ в NMDj и NMD Classifier был одинаковым, это еще раз говорит о том, что не только наилучший транскрипт-партнер, но и другие транскрипты вносят существенный вклад в значение Ψ.

Чтобы подтвердить, что предсказанные NMDj события AC действительно приводят к образованию NMDT, мы проанализировали изменения Ψ событий AC, приводящих и не приводящих к появлению NMDT, в экспериментах по инактивации системы NMD при помощи нокдауна двух ее ключевых факторов – SMG6 и SMG7 [14]. События AC, не приводящие к появлению NMDT, включали в себя кассетные экзоны, альтернативные сайты сплайсинга и удержанные интроны, найденные в кодирующих транскриптах, не являющихся NMDT. Как и ожидалось, при инактивации NMD значения Ψ событий AC, приводящих к появлению NMDT, увеличиваются более заметно, чем значения Ψ событий AC в кодирующих транскриптах (*рис. 4Б*).

ОБСУЖДЕНИЕ

Подход, реализованный в NMDj, не опирается на поиск одного наилучшего транскрипта-партнера, что позволяет ему идентифицировать и правильно описывать гораздо большее число событий AC, приводящих к появлению NMDT, по сравнению с NMD Classifier. Однако для некоторых NMDT (всего 1139 транскриптов) характеристические интроны не обнаружены, что в большинстве случаев обусловлено координацией между удаленными событиями AC и альтернативными старт- и стоп-кодонами. Например, одновременное включение и исключение несмежных экзонов 5 и 7 в гене *ERLEC1* сохраняет рамку считывания, а включение только одного экзона из пары приводит к появлению NMDT (*рис. 5А*). Этот пример показывает, что установить причинно-следственную связь между локальным событием AC и NMDT не всегда возможно, поскольку способность служить мишенью NMD является глобальной характеристикой транскрипта, которая зависит от координации между удаленными событиями AC, а локальные события AC по отдельности не определяют эти глобальные свойства. Как и любой другой подход, учитывающий только локальные события AC, NMDj принципиально не способен правильно охарактеризовать причину появления таких NMDT.

Известно, что локальные события AC регулируют экспрессию генов путем переключения AC

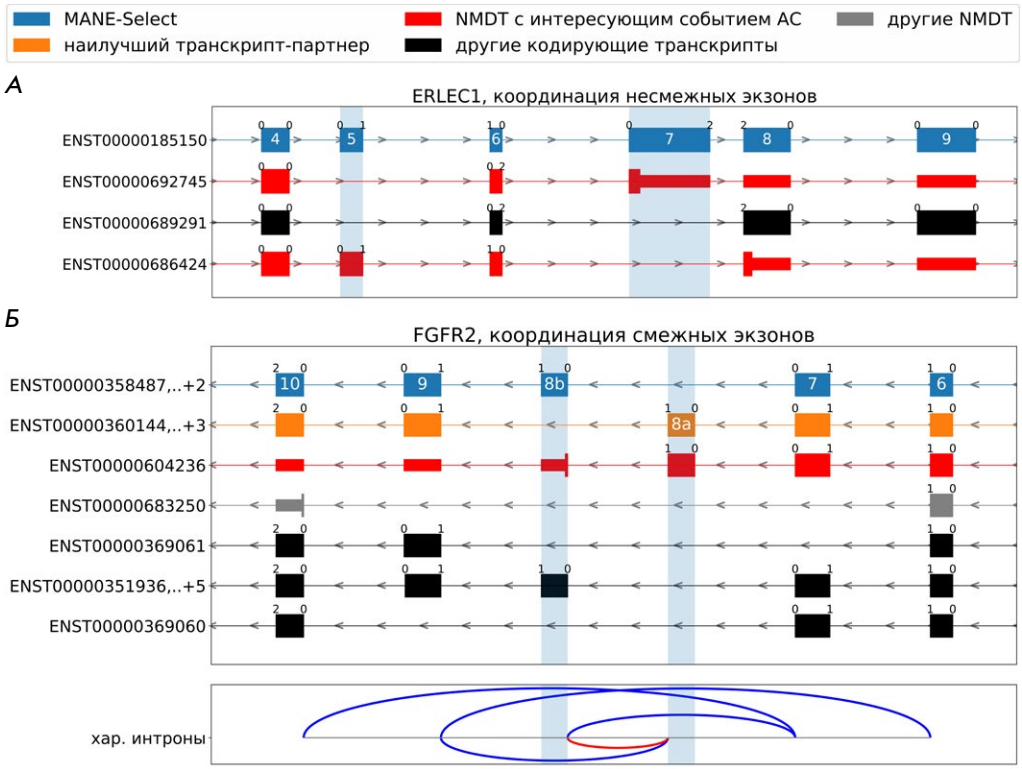

Рис. 5. Скоординированный сплайсинг несмежных кассетных экзонов в гене *ERLEC1* (*А*) и скоординированный сплайсинг экзонов в гене *FGFR2* (*Б*). Помимо взаимного исключения экзонов 8а и 8б, аннотированы изоформы с координированным пропуском экзонов 7–9 (NMDT), а также кодирующие изоформы с координированным пропуском экзонов 7–8а,б и 8а,б–9. При одновременном включении экзонов 8а и 8б возникает NMDT. Цветовые обозначения как на *рис. 2*

на производство NMDT [5, 6]. Такое переключение управляется РНК-связывающими белками, которые связываются с *цис*-элементами в пре-мРНК, и, как правило, регулируется локально [37]. С другой стороны, о функциональных последствиях и регуляторных механизмах координации событий АС на больших расстояниях известно довольно мало [38–41]. Координация между событиями АС может быть важна для получения изоформ белка с различными функциями. В некоторых случаях координация может поддерживаться системой NMD. Примером тому служит скоординированный взаимоисключающий сплайсинг экзонов 8а и 8б в гене *FGFR2*, который приводит к образованию функциональных белковых продуктов с различной лигандной специфичностью [41] (*рис. 5Б*). Включению экзона 8а способствуют эпителиально-специфичные белки ESRP1 и ESRP2, которые связываются с одной и той же регуляторной последовательностью внутри интрона [42]. При этом одновременное включение обоих экзонов приводит к образованию NMDT. Таким образом, в *FGFR2* переключение между изоформами регулируется на уровне локального АС, а координация взаимоисключения экзонов достигается за счет элиминации NMDT.

В целом, одновременный анализ всех изоформ сплайсинга вместо одного наилучшего транскрипта-партнера позволяет NMDj идентифицировать события АС, приводящие к появлению NMDT, с более высокой точностью. Однако в отношении классификации скоординированного действия удаленных событий АС NMDj имеет те же ограничения, что и другие методы. Анализ таких событий требует принципиально иных подходов. Однако представляется более вероятным, что это NMD вызывает неслучайную связь между событиями АС, а не регулируемая связь между событиями АС вызывает появление мишеней NMD. Поэтому описание скоординированных событий АС на больших расстояниях выходит за рамки данного исследования как по техническим, так и по концептуальным причинам.

Разработанный нами метод может быть полезен в исследованиях регуляции генной экспрессии по механизму непродуктивного сплайсинга [6]. Например, он может быть применен в таких задачах, как поиск специфически экспрессируемых NMDT и для оценки активности системы NMD в целом. Таким образом, NMDj закрывает существующий пробел в инструментарии для исследования сопряжения АС и NMD. ●

Представленные результаты включают данные, полученные из портала GTEx (номер dbGaP phs000424/GRU).

Данная работа выполнена при поддержке гранта Российского научного фонда (№ 22-14-00330).

Приложения доступны на сайте https://doi.org/10.32607/actanaturae.27572.

СПИСОК ЛИТЕРАТУРЫ

1. Tung K.-F., Pan C.-Y., Chen C.-H., Lin W.-C. // Sci. Rep. 2020. V. 10. № 1. P. 16245. doi: 10.1038/s41598-020-73081-5
2. Fair B., Najar C.F.B.A., Zhao J., Lozano S., Reilly A., Mossian G., Staley J.P., Wang J., Li Y.I. // Nat. Genet. 2024. V. 56. № 9. P. 1851–1861. doi: 10.1038/s41588-024-01872-x
3. Kurosaki T., Popp M.W., Maquat L.E. // Nat. Rev. Mol. Cell Biol. 2019. V. 20. № 7. P. 406–420. doi: 10.1038/s41580-019-0126-2
4. Nasif S., Contu L., Mühlemann O. // Semin. Cell Dev. Biol. 2018. V. 75. P. 78–87. doi: 10.1016/j.semcdb.2017.08.053
5. Müller-McNicoll M., Rossbach O., Hui J., Medenbach J. // J. Mol. Cell Biol. 2019. V. 11. № 10. P. 930–939. doi: 10.1093/jmcb/mjz043
6. Zavileyskiy L.G., Pervouchine D.D. // Acta Naturae. 2024. V. 16. № 1. P. 4–13. doi: 10.32607/actanaturae.27337
7. Yan Q., Weyn-Vanhentenryck S.M., Wu J., Sloan S.A., Zhang Y., Chen K., Wu J.Q., Barres B.A., Zhang C. // Proc. Natl. Acad. Sci. USA. 2015. V. 112. № 11. P. 3445–3450. doi: 10.1073/pnas.1502849112
8. Wong J.J.-L., Ritchie W., Ebner O.A., Selbach M., Wong J.W.H., Huang Y., Gao D., Pinello N., Gonzalez M., Baidya K., et al. // Cell. 2013. V. 154. № 3. P. 583–595. doi: 10.1016/j.cell.2013.06.052
9. Jangi M., Boutz P.L., Paul P., Sharp P.A. // Genes Dev. 2014. V. 28. № 6. P. 637–651. doi: 10.1101/gad.235770.113
10. Margasyuk S., Kuznetsova A., Zavileyskiy L., Vlasenok M., Skvortsov D., Pervouchine D.D. // NAR Genom. Bioinform. 2024. V. 6. № 4. P. lqae163. doi: 10.1093/nargab/lqae163
11. Mironov A., Petrova M., Margasyuk S., Vlasenok M., Mironov A.A., Skvortsov D., Pervouchine D.D. // Nucleic Acids Res. 2023. V. 51. № 7. P. 3055–3066. doi: 10.1093/nar/gkad161
12. Pervouchine D., Popov Y., Berry A., Borsari B., Frankish A., Guigó R. // Nucleic Acids Res. 2019. V. 47. № 10. P. 5293–5306. doi: 10.1093/nar/gkz193
13. Hsu M.-K., Lin H.-Y., Chen F.-C. // PLoS One. 2017. V. 12. № 4. P. e0174798. doi: 10.1371/journal.pone.0174798
14. Karousis E.D., Gypas F., Zavolan M., Mühlemann O. // Genome Biol. 2021. V. 22. № 1. P. 223. doi: 10.1186/s13059-021-02439-3
15. Isken O., Maquat L.E. // Nat. Rev. Genet. 2008. V. 9. № 9. P. 699–712. doi: 10.1038/nrg2402
16. Harrison P.W., Amode M.R., Austine-Orimoloye O., Azov A.G., Barba M., Barnes I., Becker A., Bennett R., Berry A., Bhai J., et al. // Nucleic Acids Res. 2024. V. 52. № D1. P. D891–D899. doi: 10.1093/nar/gkad1049
17. Cunningham F., Allen J.E., Allen J., Alvarez-Jarreta J., Amode M.R., Armean I.M., Austine-Orimoloye O., Azov A.G., Barnes I., Bennett R., et al. // Nucleic Acids Res. 2022. V. 50. № D1. P. D988–D995. doi: 10.1093/nar/gkab1049
18. Britto-Borges T., Gehring N.H., Boehm V., Dieterich C. // RNA. 2024. V. 30. № 10. P. 1277–1291. doi: 10.1261/rna.080066.124
19. Morales J., Pujar S., Loveland J.E., Astashyn A., Bennett R., Berry A., Cox E., Davidson C., Ermolaeva O., Farrell C.M., et al. // Nature. 2022. V. 604. № 7905. P. 310–315. doi: 10.1038/s41586-022-04558-8
20. Ivanov T.M., Pervouchine D.D. // Genes. 2018. V. 9. № 7. doi: 10.3390/genes9070356
21. Ma C., Zheng H., Kingsford C. // Algorithms Mol. Biol. 2021. V. 16. № 1. P. 5. doi: 10.1186/s13015-021-00184-7
22. Lonsdale J., Thomas J., Salvatore M., Phillips R., Lo E., Shad S., Hasz R., Walters G., Garcia F., Young N., et al. // Nat. Genet. 2013. V. 45. № 6. P. 580–585. doi: 10.1038/ng.2653
23. Yeo G., Holste D., Kreiman G., Burge C.B. // Genome Biol. 2004. V. 5. № 10. P. R74.
24. Barbosa-Morais N.L., Irimia M., Pan Q., Xiong H.Y., Gueroussov S., Lee L.J., Slobodeniuc V., Kutter C., Watt S., Colak R., et al. // Science. 2012. V. 338. № 6114. P. 1587–1593. doi: 10.1126/science.1230612
25. Li B., Dewey C.N. // BMC Bioinformatics. 2011. V. 12. P. 323. doi: 10.1186/1471-2105-12-323
26. Dobin A., Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T.R. // Bioinformatics. 2013. V. 29. № 1. P. 15–21. doi: 10.1093/bioinformatics/bts635
27. Pervouchine D.D., Knowles D.G., Guigó R. // Bioinformatics. 2013. V. 29. № 2. P. 273–274. doi: 10.1093/bioinformatics/bts678
28. Patro R., Duggal G., Love M.I., Irizarry R.A., Kingsford C. // Nat. Methods. 2017. V. 14. № 4. P. 417–419. doi: 10.1038/nmeth.4197
29. Pertea M., Pertea G.M., Antonescu C.M., Chang T.-C., Mendell J.T., Salzberg S.L. // Nat. Biotechnol. 2015. V. 33. № 3. P. 290–295. doi: 10.1038/nbt.3122
30. Popp M.W., Maquat L.E. // Cell. 2016. V. 165. № 6. P. 1319–1322. doi: 10.1016/j.cell.2016.05.053
31. Shen S., Park J.W., Lu Z., Lin L., Henry M.D., Wu Y.N., Zhou Q., Xing Y. // Proc. Natl. Acad. Sci. USA. 2014. V. 111. № 51. P. E5593–5601. doi: 10.1073/pnas.1419161111
32. Li Y.I., Knowles D.A., Humphrey J., Barbeira A.N., Dickinson S.P., Im H.K., Pritchard J.K. // Nat. Genet. 2018. V. 50. № 1. P. 151–158. doi: 10.1038/s41588-017-0004-9
33. Vaquero-Garcia J., Barrera A., Gazzara M.R., González-Vallinas J., Lahens N.F., Hogenesch J.B., Lynch K.W., Barash Y. // eLife. 2016. V. 5. P. e11752. doi: 10.7554/eLife.11752
34. Wang X., Dalkic E., Wu M., Chan C. // Curr. Opin. Biotechnol. 2008. V. 19. № 5. P. 482–491. doi: 10.1016/j.copbio.2008.07.011
35. Kim E., Magen A., Ast G. // Nucleic Acids Res. 2007. V. 35. № 1. P. 125–131.
36. Conti L.D., Baralle M., Buratti E. // Wiley Interdiscip. Rev. RNA. 2013. V. 4. № 1. P. 49–60. doi: 10.1002/wrna.1140
37. Barash Y., Calarco J.A., Gao W., Pan Q., Wang X., Shai O., Blencowe B.J., Frey B.J. // Nature. 2010. V. 465. № 7294. P. 53–59. doi: 10.1038/nature09000
38. Joglekar A., Foord C., Jarroux J., Pollard S., Tilgner H.U. // Transcription. 2023. V. 14. № 3-5. P. 92–104. doi: 10.1080/21541264.2023.2213514
39. Fededa J.P., Petrillo E., Gelfand M.S., Neverov A.D., Kadener S., Nogués G., Pelisch F., Baralle F.E., Muro A.F., Kornblihtt A.R. // Mol. Cell. 2005. V. 19. № 3. P. 393–404.
40. Bushra S., Lin Y.-N., Joudaki A., Ito M., Ohkawara B., Ohno K., Masuda A. // Int. J. Mol. Sci. 2023. V. 24. № 8. P. 7420. doi: 10.3390/ijms24087420
41. Holzmann K., Grunt T., Heinzle C., Sampl S., Steinhoff H., Reichmann N., Kleiter M., Hauck M., Marian B. // J. Nucleic Acids. 2012. V. 2012. P. 950508. doi: 10.1155/2012/950508
42. Warzecha C.C., Sato T.K., Nabet B., Hogenesch J.B., Carstens R.P. // Mol. Cell. 2009. V. 33. № 5. P. 591–601. doi: 10.1016/j.molcel.2009.01.025