МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2025, том 59, № 1, с. 22–31

ОБЗОРЫ

УДК 575.8

РОЖДЕНИЕ ГЕНОВ *de novo*

© 2025 г. Е. О. Аристова[a], И. А. Вольхин[a, b, *], А. А. Денисова[a], П. А. Никитин[a, c], Е. Р. Петрухин[a]

[a] *Факультет биоинженерии и биоинформатики Московского государственного университета им. М.В. Ломоносова, Москва, 119234 Россия*

[b] *Центр живых систем, Московский физико-технический институт (национальный исследовательский университет), Долгопрудный, 141700 Россия*

[c] *Институт проблем экологии и эволюции им. А.Н. Северцова Российской академии наук, Москва, 119071 Россия*

**e-mail: ilyavolkhin2@gmail.com*

Поступила в редакцию 25.06.2024 г.
После доработки 25.06.2024 г.
Принята к публикации 25.06.2024 г.
Рекомендована к публикации М.С. Гельфандом

Согласно классическим представлениям новые гены образуются из старых путем дупликации или горизонтального переноса. Анализ большого числа геномов, проведенный за последние десятилетия, показал, что часть генов не имеет видимых гомологов и, как предполагается, появилась *de novo* из ранее некодирующих последовательностей. В обзоре рассмотрены возможные механизмы *de novo* формирования генов, свойства кодируемых ими аминокислотных последовательностей, особенности экспрессии и отбора. Отдельно обсуждается проблема идентификации *de novo* генов.

Ключевые слова: *de novo* гены, орфанные гены, повсеместная транскрипция, повсеместная трансляция

DOI: 10.31857/S0026898425010025**, EDN:** HDFFWO

ВВЕДЕНИЕ

Ген — это транскрибируемый участок генома, кодирующий функциональный продукт: РНК или белок. Традиционно считается, что новые гены образуются из старых в ходе дупликации или горизонтального переноса. В первом случае предковая последовательность удваивается и начинает кодировать два продукта, которые затем эволюционируют независимо; во втором — фрагмент одного генома попадает в геном, где ранее не было генов, кодируемых этим фрагментом. Эти процессы лежат в основе появления в клетках РНК и белков с новыми функциями. Еще один механизм возникновения новых генов — это рекомбинация и слияние уже существующих [1].

Однако секвенирование и анализ геномов большого числа организмов показали, что многие геномы содержат гены, уникальные для вида или узкой систематической группы (рода или семейства), и не имеющие гомологов в геномах других организмов. Такие гены называются орфанными; и один из возможных механизмов их образования — возникновение *de novo*, т.е. на месте ранее некодирующей последовательности. При этом к *de novo* генам относятся как РНК-гены, так и гены, кодирующие белки. Предметом настоящего обзора являются лишь последние, в том числе потому, что возникновение *de novo* генов, кодирующих нетранслируемую РНК, почти невозможно отличить от транскрипционного шума.

Следует заметить, однако, что возможна ситуация, когда ген эволюционировал так быстро, что утратил сходство с другими известными последовательностями. В таком случае он также будет классифицирован как орфанный, хотя не является *de novo* геном [2]. Большинство исследований в данной области фокусируется на молодых *de novo* генах, т.е. наблюдаемых в ограниченном количестве таксонов, в то время как *de novo* происхождение древних генов представляет больший интерес, но крайне трудно для изучения [3]. *De novo* гены имеют ряд характер-

Сокращения: ОРС — открытая рамка считывания.

ных особенностей. Они, как правило, кодируют более короткие открытые рамки считывания (ОРС), имеют меньше экзонов и более низкий уровень экспрессии, чем другие гены [4].

Формирование *de novo* гена требует появления ОРС и регуляторных участков, необходимых для транскрипции. Есть примеры как возникновения ОРС в уже транскрибируемом участке [5], так и появления транскрипции предсуществующей ОРС (например, ген гликопротеина антифриза *AFGP*, который появился *de novo* у арктической трески [6]). Возникновение новой ОРС на месте, где ранее уже шла транскрипция, возможно как в некодирующей, так и в уже кодирующей последовательности. Во втором случае ген может считаться *de novo*, если аминокислотная последовательность его продукта никак не связана с ранее закодированной. При этом в результате сдвига ОРС или изменения сплайсинга может появиться не только полностью новый ген, но и новый участок существующего гена, однако в настоящем обзоре такие случаи мы не будем рассматривать.

Не до конца ясно, каким образом отбор поддерживает появление ОРС в нетранскрибируемых участках. Ответом на этот вопрос может быть открытие так называемой повсеместной (pervasive) транскрипции, т.е. транскрипции большей части генома, а не только участков, традиционно считаемых генами. Показано, что в клетках человека может транскрибироваться до 93% всего генома [7], а в клетках дрожжей *Saccharomyces cerevisiae*, растущих на богатой среде до 85% [8]. Возможно, благодаря этой неспецифической транскрипции случайно появившиеся ОРС могут поддерживаться отбором еще до возникновения промотора. В свою очередь, экспериментально показано, что промоторы могут формироваться случайным образом из нефункциональных участков в результате мутаций [9].

Обнаружено также явление повсеместной трансляции [10], состоящее в том, что транслироваться могут не только основные ОРС мРНК, но и другие участки мРНК, а также другие типы РНК, в частности, длинные некодирующие РНК (lncRNA). Благодаря существованию такой неспецифичной трансляции отбору может подвергаться произвольный транскрибируемый участок генома, что потенциально способно привести к появлению там *de novo* гена [11]. Механизмы формирования *de novo* генов суммированы на рис. 1.

Следует отметить, что *de novo* гены могут участвовать в процессах, принципиальных для

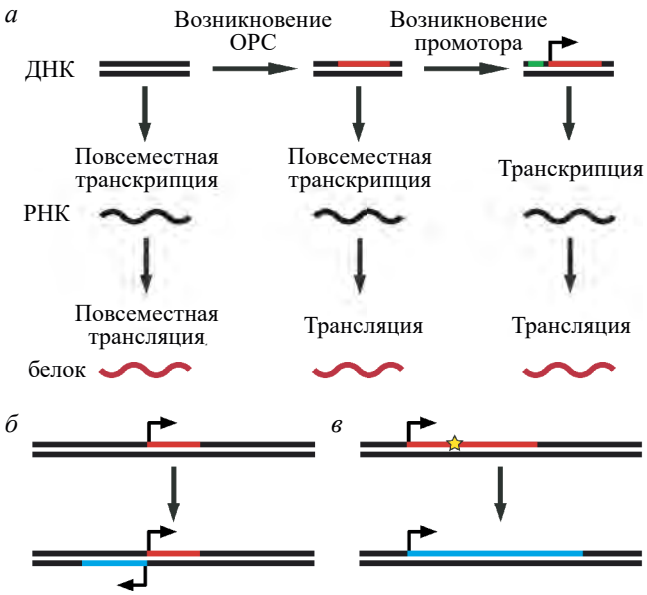

Рис. 1. Механизмы образования генов *de novo*. *а* — Возникновение генов вследствие повсеместной транскрипции и трансляции. Красным цветом обозначен *de novo* ген, зеленым – возникший промотор. *б* — Возникновение гена *de novo* на цепи, комплементарной к уже существующему гену. *в* — Возникновение гена *de novo* путем сдвига ОРС. Звездой обозначено появление точечной мутации, приводящей к сдвигу ОРС.

выживания организма, например, регулировать защитные реакции растений на патогены, как ген *OsDR10* [12], или предотвращать замерзание клеточных жидкостей в чрезвычайно холодных условиях, как белки антифризы тресковых рыб [13]. Большая межвидовая вариабельность возникших *de novo* антифризов тресковых рыб позволяет рассматривать *de novo* гены в качестве ресурса, позволяющего гибкую адаптацию к новым условиям.

Таким образом, процесс возникновения генов *de novo* представляет интерес как источник адаптации, повлиявшей на облик ныне живущих и уже исчезнувших видов, механизм которого не до конца понятен, а исследование затруднено в связи с методологическими трудностями, такими как проблема дифференциации между *de novo* и быстро эволюционирующими генами.

ВОЗНИКНОВЕНИЕ *de novo* ГЕНОВ ЧАЩЕ ПРОИСХОДИТ В РЕЗУЛЬТАТЕ ПОЯВЛЕНИЯ ОРС НА БАЗЕ УЖЕ ТРАНСКРИБИРУЕМОГО УЧАСТКА

Первым шагом при возникновении *de novo* гена в месте, где ранее не было кодирующей последовательности, может быть как приобретение транскрипции, так и появление ОРС. Последо-

вательность этих событий можно определить сравнительно-геномным анализом: если у родственных организмов какой-либо участок транскрибировался, но не содержал ОРС, можно считать, что сначала возникла транскрипция [14].

Чаще возникновение генов *de novo* происходит в результате появления ОРС внутри уже транскрибируемого участка [15–17]. Например, теломеразная РНК человека (hTERC) содержит ОРС белка, участвующего в адаптации клеток к стрессу (hTERP) [18]. Для промоторных областей *de novo* генов характерно увеличение количества предсказанных, хотя и не подтвержденных экспериментально, сайтов связывания транскрипционных факторов по сравнению с соответствующими некодирующими транскрибируемыми последовательностями близких видов [5]. Можно предположить, что *de novo* гены чаще формируются рядом с уже существующими генами и используют их промоторы, транскрипция с которых может идти в двух направлениях. Однако в этом случае *de novo* гены имели бы двунаправленные промоторы чаще, чем остальные гены, что не подтверждается анализом. Тем самым, в большинстве случаев промоторы *de novo* генов также образуются *de novo* [19]. При этом возникновение промоторов *de novo* не кажется редким событием, поскольку даже случайная последовательность может быть хорошим промотором [20].

Возможен и обратный вариант. За последние 43 млн. лет в линии, происходящей от общего предка высших приматов (Simiiformes), *de novo* появилось 19 функциональных микропротеинов, т.е. белков, длиной менее 100 аминокислот. Два таких белка человекоспецифичны, видимо, они возникли после разделения человека и шимпанзе и представляют случаи первичного появления ОРС с последующим быстрым возникновением транскрипции [14].

Образование *de novo* генов может также происходить в результате формирования новой ОРС в уже кодирующей последовательности. Это приводит к тому, что в геноме можно наблюдать перекрывание ОРС, кодирующих функциональные пептиды. Такое перекрывание характерно скорее для прокариот, особенно вирусов, имеющих ограничения на размер генома, но встречается и у эукариот. Перекрываются ОРС генов, кодирующих белки XLas и ALEX, аминокислотные последовательности которых считываются с одной мРНК, но полностью различны за счет сдвига ОРС на +1, и генов *p16Ink4a* и *p19ARF*, когда возникают два альтернативных транскрипта с разными ОРС [21].

ТРАНСКРИПТЫ МОГУТ ПОЯВЛЯТЬСЯ *de novo* НА КОМПЛЕМЕНТАРНЫХ ЦЕПЯХ УЖЕ СУЩЕСТВУЮЩИХ ГЕНОВ

Как правило, *de novo* гены возникают в межгенных областях и интронах, которые составляют большую долю генома эукариот, однако есть примеры *de novo* генов на комплементарных цепях уже существующих генов. Это показано при сравнении ОРС из транскриптомов опоссума, кенгуру, крысы, мыши и человека. Оказалось, что большинство ОРС возникают из некодирующих областей ДНК и интронных участков, но быстро теряются при дрейфе. В то же время многие ОРС перекрывались с экспрессирующимися генами на прямой и обратной цепи. Кроме того, у мыши с 14% ОРС связывались рибосомы, что может послужить признаком появления гена [22]. Другой пример — ген *NCYM*, пересекающийся по комплементарной цепи с онкогеном *MYCN*. Уровень экспрессии *NCYM* коррелирует с уровнем экспрессии *MYCN*, а его продукт участвует в регуляции онкогенеза, особенно нейробластомы, ингибируя киназу GSK3β, которая, в свою очередь, способствует деградации продукта гена *MYCN* [23].

Похожее явление обнаружено у дрожжей *S. cerevisiae*. Ген *MDF1* расположен на комплементарной цепи гена *ADF1*, представленного у всех родственных дрожжей, и подвергается положительному отбору. При этом *MDF1* не обнаружен ни у одного из представителей гемиаскомикот, кроме *S. cerevisiae*. Белок MDF1, связываясь с белком Matalpha2, играет важную роль в жизни колонии, действуя как супрессор в пути спаривания дрожжей и уменьшая тем самым эффективность их размножения. Кроме того, MDF1 участвует в регуляции метаболизма, увеличивая потребление глюкозы и скорость удвоения клеток на ранней стадии роста на богатых средах [24]. Штаммы с репрессией ADF1, но не MDF1, растут лучше клеток дикого типа, но не размножаются. ADF1 — это фактор транскрипции, который связывается с промотором антисмыслового гена *MDF1* и негативно регулирует его экспрессию. Ген *ADF1* консервативен у представителей *Saccharomyces* sensu stricto, молекулярная функция этого гена неизвестна, но его делеция, например у *S. paradoxus*, приводит к нарушению роста колоний. Кроме того, содержание ADF1 в клетке меняется в зависимости от стадии жизненного цикла [25]. Поэтому, судя по всему, ADF1 изначально был транскрипционным фактором, регулирующим процессы роста, а способность негативно влиять на экспрессию *MDF1* приобрел позже. У видов, родственных *S.*

cerevisiae, в области, гомологичной ОРС *MDF1*, расположены нетранслируемые области с многочисленными стоп-кодонами и сдвигами ОРС. Вероятно, *MDF1* появился *de novo* путем множественных мутаций в этой области [25].

De novo гены на антисмысловых цепях существующих генов дрожжей также были описаны с помощью секвенирования транскриптома [26]. Последовавший филогенетический анализ показал существование неаннотированных транскриптов, специфичных для клады, представленной *S. cerevisiae*, *S. paradoxus* и *S. mikatae*. Кроме того, показано, что уровень их экспрессии меняется вместе с ранее известными генами. Таким образом, они, вероятно, уже вовлечены в общие цепочки регуляции экспрессии. Отмечено, что заметная часть обнаруженных *de novo* транскриптов экспрессируется с уже существующих консервативных генов, но лишь с какой-то части комплементарной им цепи. С помощью рибосомного профайлинга показана трансляция части обнаруженных *de novo* транскриптов. Помимо *MDF1*, в кладе, представленной *S. cerevisiae*, *S. paradoxus* и *S. mikitae*, обнаружено возможное *de novo* происхождение нескольких генов, чья функция была известна, на комплементарной цепи уже существующих генов. Белок AUA1 участвует в транспорте аминокислот через мембрану, а кодирующий его ген *AUA1* возник напротив гена *WWM1*, участвующего в апоптозе; VAM1 участвует в образовании вакуолей, *VAM1* перекрывается по комплементарной цепи с геном, кодирующим локализованный в аппарате Гольджи мембранный белок VPS5; продукты двух транслируемых ОРС (64 и 37 кодонов), комплементарных, соответственно, двум паралогичным копиям гена *CUP1*, участвуют в снижении большой концентрации ионов меди и кадмия-44; транслируемая ОРС, кодирующая 54 аминокислотных остатка, перекрывается с геном *ARA1* арабинозодегидрогеназы.

Транскрипция с антисмысловых цепей генов широко распространена и обнаружена во всех царствах живых организмов [27]. Частое образование *de novo* генов на основе ранее существующих может быть связано с тем, что если ген уже активно экспрессируется, то велика вероятность, что комплементарная цепь тоже будет транскрибироваться в силу общей регуляции и локальной открытости хроматина. При этом комплементарный транскрипт может изначально осуществлять функцию антисмысловой регуляции [27], а затем приобретать новые функции, в частности ОРС.

de novo ГЕНЫ ОБЫЧНО ИМЕЮТ ТКАНЕСПЕЦИФИЧНУЮ ЭКСПРЕССИЮ

Экспрессия *de novo* генов часто тканеспецифична. Среди органов с высоким уровнем экспрессии *de novo* генов выделяются семенники. Считается, что из-за особенностей регуляции транскрипции и более высокой доли открытого хроматина в иммунопривилегированных семенниках создаются оптимальные условия для экспрессии негенных последовательностей, которая необходима для возникновения генов *de novo.* Показано, что уровень экспрессии *de novo* генов в семенниках сопоставим с уровнем экспрессии консервативных генов [5, 28, 29]. Тканеспецифичная экспрессия как минимум одного *de novo* гена (Shj) обнаружена в фаллопиевых трубах самок мыши в определенный период эстрального цикла [30].

De novo гены могут возникать из участков lncRNA [17], которые, как сказано выше, также могут транслироваться. Повышенная экспрессия lncRNA, характерная для семенников, но не женских репродуктивных органов [31], может объяснять высокий уровень экспрессии *de novo* генов в этом органе. Интересно, что высоким был также уровень экспрессии *de novo* генов, участвующих в эмбриональном развитии и гаметогенезе, в семенах и репродуктивных органах *Arabidopsis thaliana* [32].

Не до конца ясно, что является причиной, а что следствием: семенники имеют более высокий уровень первазивной экспрессии, поэтому в них чаще обнаруживается экспрессия *de novo* генов, или высокий уровень транскрипции различных участков генома, в том числе *de novo* генов, позволяет семенникам проявлять более широкое фенотипическое разнообразие и, тем самым, получать преимущество в ходе отбора на уровне половых клеток. Эта ситуация может быть описана в рамках концепции эволюции эволюционируемости (evolution of evolvability) [33, 34], что означало бы, что первазивная транскрипция, равно как и транскрипция *de novo* генов, сама по себе может быть основанием для отбора.

Иногда среди органов с высоким уровнем экспрессии *de novo* генов также выделяют мозг [15, 16], однако это подтверждается не во всех работах [5].

de novo ГЕНЫ ЧАЩЕ КОДИРУЮТ ТРАНСМЕМБРАННЫЕ ДОМЕНЫ

Многообразие ОРС можно рассматривать как непрерывное пространство переходных состояний от некодирующего участка через прото-

гены к устойчивому гену. Основные критерии, по которым ОРС можно считать геном, — разумная длина, признаки трансляции и очищающего отбора [35]. Проведено изучение особенностей возникающих ОРС дрожжей, т.е. молодых ОРС, экспрессия функционального продукта которых не подтверждена [36]. Предполагается, что такие ОРС являются источником *de novo* генов. Хотя повышенная экспрессия большинства возникающих ОРС была нейтральной, часть ОРС оказалась потенциально адаптивной. Предсказывалась повышенная склонность продуктов таких ОРС к формированию трансмембранных доменов, причем такой зависимости между представленностью трансмембранных доменов и адаптивностью у устоявшихся ОРС не наблюдалось. Более того, в гипотетическом продукте трансляции реконструированной предковой негенной последовательности одного из таких локусов уже предсказывается наличие участков, кодирующих трансмембранные домены, хотя и прерванные стоп-кодонами. Дальнейшая эволюция шла с сохранением склонности к формированию трансмембранного домена и удлинением ОРС. Текущий вариант локуса потенциально может кодировать небольшой интегральный мембранный белок и подвержен положительному отбору. Таким образом, наличие трансмембранных доменов в потенциальном продукте возникающей ОРС может увеличивать ее шансы закрепиться путем отбора и стать полноценным *de novo* геном. Мембранная локализация может обеспечивать защиту как самого белка от протеасомной деградации, так и организма от нежелательных взаимодействий нового объекта с содержимым цитоплазмы. После превращения протогена в полноценный *de novo* ген трансмембранная локализация может утрачиваться.

de novo ГЕНЫ МОГУТ ПОДВЕРГАТЬСЯ СТАБИЛИЗИРУЮЩЕМУ ОТБОРУ

Показано, что орфанные гены могут подвергаться стабилизирующему отбору. Так, например, у *Drosophila pseudoobscura* обнаружены 1152 орфанных гена, отсутствующих в выдаче поиска с помощью BLAST среди геномов 10 видов рода *Drosophila*, не входящих в группу *obscura* [37]. Все орфанные гены, обнаруженные одновременно у *D. pseudoobscura* и *D. affinis* (базального вида группы *obscura*), имели более высокое медианное значение dN/dS, чем старые гены, но меньшее, чем в выборке, состоящей из случайно выбранных межгенных участков, где метрика dN/dS не имеет смысла и задает фоновое распределение. По-видимому, это означает, что по крайней мере часть обнаруженных орфанных генов подвергается стабилизирующему отбору, хотя и слабому.

Однако подверженность стабилизирующему отбору может быть еще и важным критерием отличия *de novo* гена от транскрипционного шума. Так, например, у восьми представителей рода мухомор (*Amanita*) обнаружено 109 семейств орфанных генов; далее проверяли отсутствие кодирующих последовательностей в соответствующих участках генома родственных видов, что может говорить об их возможном *de novo* происхождении [38]. В результате два семейства генов классифицировали как вероятно появившиеся *de novo* у некоторых представителей рода. У каждого из этих семейств значения $dN/dS << 1$, что позволяет предположить действие стабилизирующего отбора, а их GC-состав оказался близок по значению к GC-составу старых генов. Функциональную значимость этих генов выявить не удалось.

АНАЛОГИ *de novo* ГЕНОВ ЕСТЬ И У ПРОКАРИОТ

Геномы прокариот гораздо более компактны, чем у эукариот, они содержат меньше некодирующих участков, так что можно ожидать, что *de novo* гены у них не возникают или это гораздо менее вероятное событие. Тем не менее в геномах прокариот также находятся неконсервативные и видоспецифические последовательности, так называемые мОРС (малые ОРС, менее 100 кодонов), предположительно возникающие из некодирующих участков *de novo* [39]. Экспрессирующиеся с мОРС белки, функции которых установлены, участвуют в большом количестве клеточных процессов: от морфогенеза и клеточного деления бактерий до транспорта и реакций на стресс. Среди них можно выделить белки SpoVM и CmpA, участвующие в споруляции; MciZ, SidA и Blr, участвующие в делении клеток; регуляторы транспорта KdpF, AcrZ и SgrT, взаимодействующие с транспортерами; регуляторы мембраносвязанных ферментов CydX, PmrR и MgtR [40].

В геноме *Escherichia coli* предсказаны 125 мОРС, с которых экспрессируются пептиды длиной менее 72 аминокислотных остатков [41]. Может ли возникновение случайных коротких пептидов поддерживаться отбором? Ответу на этот вопрос посвящены экспериментальные исследования, в одном из которых использовали ауксотрофный штамм *E. coli* с мутацией в гене *serB*, кодирующем фосфатазу, катализирующую

превращение 3-фосфосерина в серин [39]. Из искусственной библиотеки генов случайных коротких пептидов смогли выбрать такие, которые позволили ауксотрофному штамму выживать на среде без серина. Оказалось, что это обусловлено связыванием этих пептидов с мРНК другой фосфатазы — HisB, которое приводит к повышенной экспрессии соответствующего белка, способного заменять SerB. Еще в одной работе в клетках *E. coli* экспрессировали более 100 млн. случайно сгенерированных последовательностей ДНК и отобрали шесть вариантов, которые кодируют пептиды, обеспечивающие устойчивость к антибиотику колистину [42]. Таким образом, экспериментально показано, что возникновение *de novo* генов у прокариот может быть поддержано отбором.

de novo ГЕНЫ СЛОЖНО ОТЛИЧИТЬ ОТ БЫСТРО ЭВОЛЮЦИОНИРУЮЩИХ ГЕНОВ

Как уже сказано, орфанные гены — это гены, уникальные для вида или небольшой таксономической группы, не имеющие идентифицируемых гомологов в геномах других организмов. Существуют два гипотетических механизма возникновения орфанных генов. Первый — возникновение из некодирующей последовательности, механизм *de novo*. Второй — возникновение в результате дупликации или видообразования, за которыми следует быстрая дивергенция, в результате чего ген теряет сходство с гомологами. На практике эти два механизма очень сложно различить.

Основным инструментом для определения гомологов является BLAST, который находит области локального сходства между последовательностями [43]. Как правило, ген считается орфанным, если с помощью алгоритма BLAST не удалось выявить его гомологи. На практике результаты зависят от используемого варианта BLAST (BLASTP, BLASTX, BLASTN) и порогового сходства (измеряемого как E-value, количество последовательностей с данным уровнем сходства в выборке случайных последовательностей того же объема, как и GenBank). Это делает идентификацию орфанных генов достаточно субъективной. Алгоритм BLAST сам по себе не отвечает на вопрос о механизме происхождения гена [44, 45]; он предназначен для обнаружения сходства последовательностей ДНК или белков, а не гомологии, т.е. общего происхождения, но гомология не всегда означает обнаруживаемое сходство, поскольку гомологи с сильно дивергировавшими последовательностями могут быть пропущены BLAST. В этих случаях неспособность обнаружить гомологи отражает предел разрешающей способности алгоритма BLAST, а не биологическую реальность. Это называется филостратиграфической ошибкой [46]. На обнаружение гомологов гена с помощью BLAST влияет также его длина и длина ОРС, так как короткие и быстро эволюционирующие белки можно не заметить даже у близких видов, при том что они реально существуют [47].

Возможность появления орфанных генов в ходе дупликации с последующей быстрой эволюцией и потерей видимой гомологии широко обсуждается (рис. 2). На сегодняшний момент отсутствует полное понимание механизма эволюции дуплицирующихся генов. Почему после дупликации сохраняются обе копии гена и почему их эволюционная судьба различна? Считается, что эволюционная судьба копий гена определяется природой мутаций, которые возникают в двух копиях, обеспечивающих основу для отбора. Паралоги сохраняются с помощью различных механизмов, включая неофункционализацию (возникновение новых функций у одной из копий) [48–50], разделение одной функции между двумя генами путем субфункционализации [51–53], дозированную амплификацию одной из копий [54].

Таким образом, стандартный метод поиска орфанных генов *de novo* не позволяет отличить такие гены от быстро эволюционирующих. Однако, используя дополнительные соображения, иногда удается показать, что орфанный ген возник именно из некодирующей последовательности, либо просто отбросить вариант с возникновением гена в ходе дупликации с последующей дивергенцией.

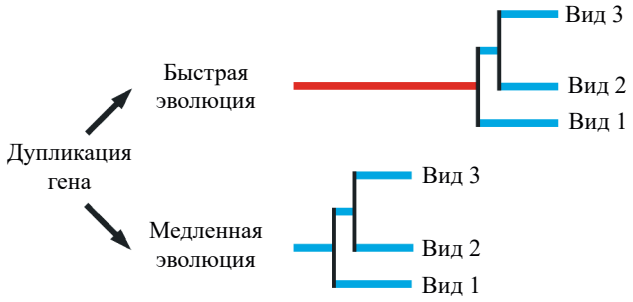

Рис. 2. Быстрая эволюция дуплицированного гена может привести к накоплению значительного количества замен, вследствие чего он не будет идентифицироваться как гомолог других последовательностей с помощью BLAST [43] и, следовательно, будет признан орфанным.

Как упомянуто выше, один из основных методов доказательства происхождения гена *de novo* — поиск в родственных группах организмов гомологичной геномной области, которая является некодирующим участком, не подвергающимся транскрипции и трансляции.

Наиболее доступным способом выявления таких областей считается построение блоков синтении. Блоки синтении — это протяженные подпоследовательности, не испытавшие перестроек, и обнаруживаемые при сравнении двух и более хромосом. Построение синтенных блоков позволяет найти гомологичные участки, в которых и следует искать некодирующего предка *de novo* гена. С использованием этого подхода у представителей рода мухомор (*Amanita*) были выделены два семейства генов, описанных как вероятно появившиеся *de novo* [38]. Этот метод доказательства использовали также при изучении гена *FLJ33706*, который экспрессируется в мозге человека и предположительно связан с никотиновой зависимостью и болезнью Альцгеймера [55]. *FLJ33706* произошел из некодирующих последовательностей ДНК: вставка повторяющихся элементов, особенно Alu, способствовала образованию первого кодирующего экзона и шести пар сайтов, задающих интроны, на ветви, ведущей к человеку и шимпанзе, а две последующие замены в человеческой линии позволили избежать двух стоп-кодонов, которые не позволяют транслироваться соответствующей ОРС у шимпанзе.

Исходя из максимальной скорости замен в данной группе организмов, иногда удается показать, что быстрая потеря идентифицируемого сходства между гомологами маловероятна. С помощью этих соображений показано, что ген *OsDR10,* специфичный для риса (*Oryza)*, скорее возник *de novo*, чем утратил наблюдаемую гомологию в ходе накопления замен [12].

Таким образом, с помощью BLAST можно идентифицировать орфанные гены, но для доказательства возникновения гена *de novo* необходимо привлекать дополнительные аргументы. Наилучшим способом, указывающим на возможное *de novo* происхождение гена, является построение синтенных блоков геномов нескольких организмов. В настоящее время разрабатываются программные подходы, которые комбинируют поиск гомологичных последовательностей с помощью выравнивания и построения блоков синтении [56]. Можно ожидать, что такие методы позволят найти другие примеры *de novo* генов.

ЗАКЛЮЧЕНИЕ

De novo рождение генов — это процесс, не вполне согласующийся с простыми представлениями об эволюции геномов. В ходе этого процесса исходно нефункциональные локусы могут начать кодировать белковые продукты и обрести полезные функции. При этом более вероятно появление *de novo* генов в областях с уже доступным хроматином, так как при этом не требуется его ремоделирование в уже активных генах или их частях, например, на комплементарной цепи, из интронов и других транскрибируемых элементов генома. Регуляция *de novo* генов часто связана с ранее существовавшими генами, поскольку они располагаются в уже готовом окружении из сайтов связывания транскрипционных факторов и открытого хроматина.

Существуют технические ограничения для поиска генов, возникших в геноме *de novo*. Поиск гомологов с помощью BLAST является легкодоступным методом, но с его использованием не удается получить исчерпывающие результаты. Помимо отсутствия гомологов необходимо также привлекать данные по экспрессии генов (например RNA-seq, Саузерн-блотинг или ПЦР), показывать наличие трансляции (с помощью рибосомного профайлинга, вестерн-блотинга или масс-спектрометрии) и отбора (например, с помощью метрики dN/dS). Кроме того, предполагается, что отсутствие видимого сходства может наблюдаться не только в случае *de novo* образования гена, но и в случае быстрой эволюции локуса. В связи с этим истинную частоту появления *de novo* генов сложно оценить.

Некоторые появившиеся *de novo* гены обладают важной функцией, например, участвуют в регуляции полового процесса или влияют на выживаемость, однако бóльшая часть *de novo* генов кодирует короткие пептиды или белки с неизвестной функцией, а их последовательности не имеют признаков очищающего отбора. Определение функций *de novo* генов затруднено тем, что у таких генов нет гомологов.

Таким образом, *de novo* рождение генов можно рассматривать в качестве важного, но до сих пор недостаточно исследованного механизма обретения организмами новых функций.

Написание обзора не потребовало специального финансирования.

Работа выполнена без привлечения людей и животных в качестве объектов исследования.

Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. Arroyo J.I., Nery M.F. (2018) Gene fusion of heterophyletic gamma-globin genes in platyrrhine primates. *J. Genet.* **97**, 1473–1478.
2. Tautz D., Domazet-Lošo T. (2011) The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702.
3. Van Oss S.B., Carvunis A.-R. (2019) *De novo* gene birth. *PLoS Genet.* **15**, e1008160.
4. Черезов Р.О., Воронцова Ю.Е., Симонова О.Б. (2021) Феномен эволюционной “генерации De Novo” генов. *Онтогенез.* **52**, 441–452.
5. Ruiz-Orera J., Hernandez-Rodriguez J., Chiva C., Sabidó E., Kondova I., Bontrop R., Marqués-Bonet T., Albà M.M. (2015) Origins of *de novo* genes in human and chimpanzee. *PLoS Genet.* **11**, e1005721.
6. Zhuang X., Yang C., Murphy K.R., Cheng C.-C. (2019) Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc. Natl. Acad. Sci. USA.* **116**, 4400–4405.
7. Clark M.B., Amaral P.P., Schlesinger F.J., Dinger M.E., Taft R.J., Rinn J.L., Ponting C.P., Stadler P.F., Morris K.V., Morillon A., Rozowsky J.S., Gerstein M.B., Wahlestedt C., Hayashizaki Y., Carninci P., Gingeras T.R., Mattick J.S. (2011) The reality of pervasive transcription. *PLoS Biol.* **9**, e1000625.
8. David L., Huber W., Granovskaia M., Toedling J., Palm C.J., Bofkin L., Jones T., Davis R.W., Steinmetz L.M. (2006) A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. USA.* **103**, 5320–5325.
9. Yona A.H., Alm E.J., Gore J. (2018) Random sequences rapidly evolve into *de novo* promoters. *Nat. Commun.* **9**, 1530.
10. Ingolia N.T., Brar G.A., Stern-Ginossar N., Harris M.S., Talhouarne G.J.S., Jackson S.E., Wills M.R., Weissman J.S. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **8**, 1365–1379.
11. Parikh S.B., Houghton C., Van Oss S.B., Wacholder A., Carvunis A. (2022) Origins, evolution, and physiological implications of *de novo* genes in yeast. *Yeast.* **39**, 471–481.
12. Xiao W., Liu H., Li Y., Li X., Xu C., Long M., Wang S. (2009) A rice gene of *de novo* origin negatively regulates pathogen-induced defense response. *PLoS One.* **4**, e4603.
13. Baalsrud H.T., Tørresen O.K., Solbakken M.H., Salzburger W., Hanel R., Jakobsen K.S., Jentoft S. (2018) *De novo* gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol. Biol. Evol.* **35**, 593–606.
14. Vakirlis N., Vance Z., Duggan K.M., McLysaght A. (2022) *De novo* birth of functional microproteins in the human lineage. *Cell Rep.* **41**, 111808.
15. Xie C., Zhang Y.E., Chen J.-Y., Liu C.-J., Zhou W.-Z., Li Y., Zhang M., Zhang R., Wei L., Li C.-Y. (2012) Hominoid-specific *de novo* protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* **8**, e1002942.
16. Wu D.-D., Irwin D.M., Zhang Y.-P. (2011) *De novo* origin of human protein-coding genes. *PLoS Genet.* **7**, e1002379.
17. An N.A., Zhang J., Mo F., Luan X., Tian L., Shen Q.S., Li X., Li C., Zhou F., Zhang B., Ji M., Qi J., Zhou W.Z., Ding W., Chen J.Y., Yu J., Zhang L., Shu S., Hu B., Li C.Y. (2023) *De novo* genes with an lncRNA origin encode unique human brain developmental functionality. *Nat. Ecol. Evol.* **7(2)**, 264–278.
18. Rubtsova M., Naraykina Y., Vasilkova D., Meerson M., Zvereva M., Prassolov V., Lazarev V., Manuvera V., Kovalchuk S., Anikanov N., Butenko I., Pobeguts O., Govorun V., Dontsova O. (2018) Protein encoded in human telomerase RNA is involved in cell protective pathways. *Nucl. Acids Res.* **46**, 8966–8977.
19. Wu X., Sharp P.A. (2013) Divergent transcription: a driving force for new gene origination? *Cell.* **155**, 990–996.
20. Vaishnav E.D., De Boer C.G., Molinet J., Yassour M., Fan L., Adiconis X., Thompson D.A., Levin J.Z., Cubillos F.A., Regev A. (2022) The evolution, evolvability and engineering of gene regulatory DNA. *Nature.* **603**, 455–463.
21. Chung W.-Y., Wadhawan S., Szklarczyk R., Pond S.K., Nekrutenko A. (2007) A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput. Biol.* **3**, e91.
22. Schmitz J.F., Ullrich K.K., Bornberg-Bauer E. (2018) Incipient *de novo* genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* **2**, 1626–1632.
23. Suenaga Y., Islam S.M., Alagu J., Kaneko Y., Kato M., Tanaka Y., Kawana H., Hossain S., Matsumoto D., Yamamoto M., Shoji W., Itami M., Shibata T., Nakamura Y., Ohira M., Haraguchi S., Takatori A., Nakagawara A. (2014) NCYM, a cis-antisense gene of *MYCN*, encodes a *de novo* evolved protein that inhibits GSK3β resulting in the stabilization of MYCN in human neuroblastomas. *PLoS Genet.* **10**, e1003996.
24. Li D., Yan Z., Lu L., Jiang H., Wang W. (2014) Pleiotropy of the *de novo*-originated gene *MDF1*. *Sci. Rep.* **4**, 7280.
25. Li D., Dong Y., Jiang Y., Jiang H., Cai J., Wang W. (2010) A *de novo* originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.* **20**, 408–420.
26. Blevins W.R., Ruiz-Orera J., Messeguer X., Blasco-Moreno B., Villanueva-Cañas J.L., Espinar L., Díez J., Carey L.B., Albà M.M. (2021) Uncovering *de novo* gene birth in yeast using deep transcriptomics. *Nat. Commun.* **12**, 604.
27. Pelechano V., Steinmetz L.M. (2013) Gene regulation by antisense transcription. *Nat. Rev. Genet.* **14**, 880–893.
28. Begun D.J., Lindfors H.A., Kern A.D., Jones C.D. (2007) Evidence for *de novo* evolution of testis-

expressed genes in the *Drosophila yakuba/Drosophila erecta* clade. *Genetics.* **176**, 1131–1137.

29. Levine M.T., Jones C.D., Kern A.D., Lindfors H.A., Begun D.J. (2006) Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. USA.* **103**, 9935–9939.
30. Xie C., Bekpen C., Künzel S., Keshavarz M., Krebs-Wheaton R., Skrabar N., Ullrich K.K., Tautz D. (2019) A *de novo* evolved gene in the house mouse regulates female pregnancy cycles. *eLife.* **8**, e44392.
31. Lombardo K.D., Sheehy H.K., Cridland J.M., Begun D.J. (2023) Identifying candidate de novo genes expressed in the somatic female reproductive tract of *Drosophila melanogaster*. *G3 Genes Genomes Genet.* **13**, jkad122.
32. Donoghue M.T., Keshavaiah C., Swamidatta S.H., Spillane C. (2011) Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evol. Biol.* **11**, 47.
33. Kirschner M., Gerhart J. (1998) Evolvability. *Proc. Natl. Acad. Sci. USA.* **95**, 8420–8427.
34. Koonin E.V., Wolf Y.I. (2012) Evolution of microbes and viruses: a paradigm shift in evolutionary biology? *Front Cell Infect. Microbiol.* **2**, 119.
35. Carvunis A.R., Rolland T., Wapinski I., Calderwood M.A., Yildirim M.A., Simonis N., Charloteaux B., Hidalgo C.A., Barbette J., Santhanam B., Brar G.A., Weissman J.S., Regev A., Thierry-Mieg N., Cusick M.E., Vidal M. (2012) Proto-genes and *de novo* gene birth. *Nature.* **487**, 370–374.
36. Vakirlis N., Acar O., Hsu B., Castilho Coelho N., Van Oss S.B., Wacholder A., Medetgul-Ernar K., Bowman R.W. 2nd, Hines C.P., Iannotta J., Parikh S.B., McLysaght A., Camacho C.J., O'Donnell A.F., Ideker T., Carvunis A.R. (2020) *De novo* emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* **11**, 781.
37. Palmieri N., Kosiol C., Schlötterer C. (2014) The life cycle of *Drosophila* orphan genes. *eLife.* **3**, e01311.
38. Wang Y.-W., Hess J., Slot J.C., Pringle A. (2020) *De novo* gene birth, horizontal gene transfer, and gene duplication as sources of new gene families associated with the origin of symbiosis in *Amanita*. *Genome Biol. Evol.* **12**, 2168–2182.
39. Babina A.M., Surkov S., Ye W., Jerlström-Hultqvist J., Larsson M., Holmqvist E., Jemth P., Andersson D.I., Knopp M. (2023) Rescue of *Escherichia coli* auxotrophy by *de novo* small proteins. *eLife.* **12**, e78299.
40. Storz G., Wolf Y.I., Ramamurthi K.S. (2014) Small proteins can no longer be ignored. *Annu. Rev. Biochem.* **83**, 753–777.
41. Knopp M., Gudmundsdottir J.S., Nilsson T., König F., Warsi O., Rajer F., Ädelroth P., Andersson D.I. (2019) *De novo* emergence of peptides that confer antibiotic resistance. *mBio.* **10**, e00837-19.
42. Knopp M., Babina A.M., Gudmundsdóttir J.S., Douglass M.V., Trent M.S., Andersson D.I. (2021) A novel type of colistin resistance genes selected from random sequence space. *PLoS Genet.* **17**, e1009227.
43. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
44. Albà M.M., Castresana J. (2007) On homology searches by protein BLAST and the characterization of the age of genes. *BMC Evol. Biol.* **7**, 53.
45. Elhaik E., Sabath N., Graur D. (2006) The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. *Mol. Biol. Evol.* **23**, 1–3.
46. Moyers B.A., Zhang J. (2017) Further simulations and analyses demonstrate open problems of phylostratigraphy. *Genome Biol. Evol.* **9**, 1519–1527.
47. Weisman C.M., Murray A.W., Eddy S.R. (2020) Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol.* **18**, e3000862.
48. Chanderbali A.S., Berger B.A., Howarth D.G., Soltis D.E., Soltis P.S. (2017) Evolution of floral diversity: genomics, genes and *gamma*. *Philos. Trans. R Soc. B Biol. Sci.* **372**, 20150509.
49. Assis R., Bachtrog D. (2013) Neofunctionalization of young duplicate genes in *Drosophila*. *Proc. Natl. Acad. Sci. USA.* **110**, 17409–17414.
50. Ohno S. (1970) *Evolution by Gene Duplication*. Heidelberg: Springer Berlin.
51. Force A., Lynch M., Pickett F.B., Amores A., Yan Y., Postlethwait J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics.* **151**, 1531–1545.
52. Lynch M., Force A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics.* **154**, 459–473.
53. Hardison R.C. (2012) Evolution of hemoglobin and its genes. *Cold Spring Harb. Perspect. Med.* **2**, a011627–a011627.
54. Kondrashov F.A., Kondrashov A.S. (2006) Role of selection in fixation of gene duplications. *J. Theor. Biol.* **239**, 141–151.
55. Li C.Y., Zhang Y., Wang Z., Zhang Y., Cao C., Zhang P.W., Lu S.J., Li X.M., Yu Q., Zheng X., Du Q., Uhl G.R., Liu Q.R., Wei L. (2010) A human-specific *de novo* protein-coding gene associated with human brain functions. *PLoS Comput. Biol.* **6**, e1000734.
56. Vakirlis N., McLysaght A. (2019) Computational prediction of *de novo* emerged protein-coding genes. *Meth. Mol. Biol.* **1851**, 63–81.

The Birth of *de novo* Genes

© 2025 E. O. Aristova[1], I. A. Volkhin[1, 2, *], A. A. Denisova[1], P. A. Nikitin[1,3], E. R. Petrukhin[1]

[1]*Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, 119234 Russia*
[2]*Life Sciences Research Center, Moscow Institute of Physics and Technology (National Research University), Dolgoprudny, Moscow Region, 141700 Russia*
[3]*Institute of Ecology and Evolution Problems named after A.N. Severtsov, Russian Academy of Sciences, Moscow, 119071 Russia*
**e-mail: ilyavolkhin2@gmail.com*

According to classical ideas, new genes emerge from old genes by duplication or horizontal transfer. Analysis of a large number of genomes in recent decades has shown that some genes have no visible homologs and are thought to have emerged *de novo* from previously noncoding sequences. The review considers possible mechanisms of *de novo* gene formation, properties of protein sequences encoded by them, features of expression and selection. The problem of *de novo* gene identification is considered separately.

Keywords: *de novo* gene, orphan gene, pervasive transcription, pervasive translation