ОНТОГЕНЕЗ, 2021, том 52, № 6, с. 441–452

ОБЗОРЫ

УДК 575.8

ФЕНОМЕН ЭВОЛЮЦИОННОГО “САМОЗАРОЖДЕНИЯ” ГЕНОВ

© 2021 г. Р. О. Черезов[a], Ю. Е. Воронцова[a], О. Б. Симонова[a], *

[a] Институт биологии развития им. Н.К. Кольцова РАН, ул. Вавилова, 26, Москва, 119334 Россия

*e-mail: osimonova@hotmail.com

Поступила в редакцию 18.06.2021 г.
После доработки 20.07.2021 г.
Принята к публикации 25.07.2021 г.

Вопрос возникновения и эволюции новых генов всегда интересовал эволюционных генетиков. Наиболее очевидными механизмами их формирования являются разного рода хромосомные и межгенные перестройки, подразумевающие использование в качестве исходного материала уже существующие гены. Возможность происхождения полностью функционального гена из некодирующей ДНК, т.е. *de novo*, не отвергалась, но до последнего времени практически сводилась к нулю. Тем не менее, в 1996 г. после анализа генома дрожжей *Saccharomyces cerevisiae* были получены первые экспериментальные доказательства возможности формирования генов *de novo*. Спустя 10 лет, гены, не имеющие гомологов, предполагаемо возникшие *de novo*, были найдены у дрозофилы. Относительно высокая вероятность возникновения генов *de novo*, оцененная в биоинформационных исследованиях, подняла интерес к этой теме и сделала актуальным их поиск. Сейчас количество работ, посвященных проблеме возникновения генов *de novo* у разных организмов, включая человека, постоянно растет, демистифицируя этот феномен. Тем не менее, остается много вопросов, требующих теоретического и практического исследования. Данный обзор посвящен проблеме поиска и характеристике генов, возникших *de novo*, а также предполагаемым механизмам их возникновения.

Ключевые слова: ген, эволюция, *de novo*

DOI: 10.31857/S0475145021060033

Вдали от равновесия могут возникать новые процессы самоорганизации.
И. Пригожин, И. Стенгерс
“Порядок из хаоса”

ВВЕДЕНИЕ

Долгое время основным механизмом формирования новых генов в эволюции считались дупликации существующих генов у предковых организмов с последующей дивергенцией возникших копий (Haldane, 1933; Muller, 1935; Ohno et al., 1968; Ohno, 1970). Зарождение генов *de novo* внутри некодирующих районов ДНК, то есть “с нуля”, считали маловероятным событием (Jacob, 1977). Тем не менее, такие предположения высказывались в работах Граса и Оно (Grasse, 1977; Ohno, 1984). В 1996 году были получены первые экспериментальные доказательства возможности формирования генов *de novo* (Dujon, 1996). В этой работе был проведен анализ генома дрожжей *Saccharomyces cerevisiae*, в результате которого были выявлены гены, не имеющие гомологов у других видов. Затем появилась серия работ, выполненных на разных видах плодовой мухи *Drosophila* (Levine et al., 2006; Begun et al., 2006, 2007). В работе Бегана (Begun et al., 2006) был проанализирован транскриптом придаточных желез репродуктивной системы самцов дрозофил видов *D. yakuba* и *D. erecta*. В результате обнаружили 20 видоспецифичных генов, которые вероятнее всего возникли *de novo*, без участия дупликаций, или других перестроек предковых последовательностей. У *D. melanogaster* и *D. simulans* было найдено 5 таких генов (Levine et al., 2006). Затем при анализе библиотеки кДНК, полученной из семенников *D. erecta* и *D. yakuba* были обнаружены еще 7 уникальных генов, возникших *de novo* (Begun et al., 2007). Параллельно с этим у *S. cerevisiae* было найдено 99 генов, специфичных только для этого вида и, вероятно, возникших *de novo* (Nishida, 2006). Также у *S. cerevisae* в 2008 г. был обнаружен возникший *de novo* ген *BSC4*, кодирующий белок не имеющий гомологов даже у близкородственных видов (Cai et al., 2008). Авторы предположили, что новый ген может являться компонентом сигнального пути репарации ДНК. В 2009 г. у *Arabidopsis thaliana* был охарактеризован первый ген (*QQS*), возникший *de novo* и предположительно участвующий в метаболизме крахмала (Li et al., 2009). У домовой мыши в 2009 г. был описан ген *Poldi,* возникший *de novo* и кодирующий только длинную некодирующую РНК (lncRNA). По дан-

ным авторов, этот ген возник в межгенном районе, который присутствует в геноме многих млекопитающих, включая человека (Heinen et al., 2009). Тем не менее транскрипты *Poldi* детектируются только в семенниках рода *Mus musculus*, на основании чего авторы заключили, что этот ген возник *de novo*, а не в результате геномных перестроек. В 2009 г. у человека были обнаружены три гена (*CLLU1*, *C22orf45*, *DNAH10OS*), возникшие уже после разделения предков современных шимпанзе и *Homo sapiens* (Knowles, McLysaght, 2009). Один из этих генов (*CLLU1*) предположительно связан с развитием хронического лимфолейкоза. Позже, поиск функциональных генов, возникших *de novo* у человека, выявил ген *FLJ33706* предположительно связанный с развитием никотиновой зависимости и болезнью Альцгеймера (Li et al., 2010).

Сейчас количество работ, посвященных проблеме возникновения генов *de novo* у разных организмов, постоянно растет. Несмотря на это, в отношении многих примеров остается открытым вопрос о том, являются ли эти гены действительно возникшими *de novo*. Тем не менее, учитывая все имеющиеся данные, можно с уверенностью сказать: возникновение *de novo* – является самостоятельным реально существующим механизмом формирования новых генов в ходе эволюции (Neme, Tautz, 2013; Oss, Carvunis, 2019).

Следует отметить, что в настоящее время в список генов, возникших *de novo*, часто включают гены-сироты (orphan genes) (Schlötterer, 2015; Zhang et al., 2019). Однако, гены-сироты – это более широкое понятие, объединяющее таксон-специфические гены, которые являются уникальными для определенного таксономического уровня. Механизмами возникновения генов-сирот могут быть: дупликации предкового гена и быстрая дивергенция его копий, в результате которой становится невозможным обнаружить ее гомологов (Tautz, Domazet-Lošo, 2011); экзаптация мобильных элементов (Toll-Riera et al., 2009); горизонтальный перенос генов (Husnik, McCutcheon, 2018); сдвиг кодирующей белок рамки считывания, приводящий к возникновению совершенно нового белка (Neme, Tautz, 2013) и собственно возникновение генов *de novo* в ранее некодирующих участках генома (Schlötterer, 2015). Таким образом, именно гены, которые возникли *de novo*, входят в группу генов-сирот, а не наоборот. Этим генам и будет посвящен данный обзор.

МЕТОДЫ ОБНАРУЖЕНИЯ ГЕНОВ, ВОЗНИКШИХ *DE NOVO*

Для поиска генов, возникших *de novo*, используют геномную филостратиграфию (Domazet-Loso et al., 2007) и подходы, в основе которых лежит синтения генов.

В геномной филостратиграфии сравниваются последовательности гомологов исследуемого гена у близкородственных и отдаленных видов, полученные с помощью программного обеспечения BLAST или его аналогов (Altschul et al., 1990). В результате этого анализа становится возможным определить возраст гена, т.е. участок на филогенетическом древе, соответствующий появлению общего предка для видов, у которых обнаруживаются гомологи изучаемого гена. Если же в ходе филостратиграфического анализа его гомологи у близкородственных видов не обнаруживаются, то он считается новым или таксон-специфичным, в зависимости от того, в какой группе видов проводили поиск. Использование филостратиграфического анализа для поиска генов, возникших *de novo*, ограничено имеющимися данными о геномах, родственных исследуемому, и используемыми критериями поиска в BLAST (McLysaght, Hurst, 2016). Так, например, часто нельзя исключить того, что исследуемый ген возник после дупликации предкового гена и последующей быстрой дивергенции его копии, не определяемой как гомолог при использовании поиска в BLAST. На обнаружение гомологов гена также влияет его длина и длина открытой рамки считывания, так как короткие и быстро эволюционирующие белки могут быть не обнаружены у родственных видов, даже если они реально существуют (Elhaik, 2006; Moyers, Zhang, 2015; Weisman, 2020). Поэтому для более успешного обнаружения филогенетически древних гомологов используются более чувствительные алгоритмы поиска такие, как CS-BLAST, PSI-BLAST (Altschul et al., 1997) и скрытые модели Маркова (Potter et al., 2018). Например, при поиске генов, возникших *de novo* методом геномной филостратиграфии было обнаружено 25 новых генов у *Saccharomyces cerevisiae* (Carvunis et al., 2012), 781 ген у *Mus musculus* (Neme, Tautz, 2013), 84 гена у *Mus musculus* (Wilson et al., 2017). Однако в других исследованиях, при моделировании эволюции геномов нескольких видов дрожжей *Saccharomyces* и мух *Drosophila* было показано, что, филостратиграфический подход не смог обнаружить ортологи в родственных видах для 11% из 5878 моделей генов дрожжей и для 13.9% из 6695 моделей генов мух, и эти гены ошибочно отнесли к генам, возникшим *de novo* (Moyers, Zhang, 2015, 2016).

В сравнении с геномной филостратиграфией, более достоверные данные при поиске генов, возникших *de novo*, можно получить используя подходы, основанные на анализе консервативных синтенных участков (синтенных блоков), так как они позволяют обнаружить в геномах видов аутгрупп участки, на основе которых возникли предполагаемые новые гены (Chen, 2010; Tautz, Domazet-Lošo, 2011; McLysaght, Hurst, 2016). Выравнивания синтенных последовательностей

обычно делают относительно выбранных консервативных маркеров: генов, *k*-меров и, иногда, экзонов (Ghiurcutam, Moret 2014; Gehrmann, Reinders, 2015). Отсутствие реальной экспрессии белка или предполагаемой рамки считывания в синтенном районе в аутгруппе позволяет с высокой степенью уверенности утверждать, что исследуемый ген действительно возник *de novo* (McLysaght, Hurst, 2016). Наиболее убедительным доказательством происхождения гена *de novo* является обнаружение специфической мутации, которая привела к трансформации ранее некодирующего участка генома в кодирующий. Такие данные обычно получают в ходе анализа микросинтенных районов у близкородственных видов (Oss, Carvunis, 2019). Одним из препятствий к использованию методов, основанных на анализе синтенных блоков, является сложность выявления синтении в филогенетически значительно удаленных друг от друга геномах. Также сложности возникают, если исследуемые геномы фрагментированы (Liu et al., 2018) или в исследуемой эволюционной линии наблюдается высокая частота спонтанных мутаций (Ranz et al., 2001). С помощью анализа консервативных синтенных участков были обнаружены возникшие *de novo* гены человека: *FLJ33706* (Li et al., 2010), *CLLU1*, *C22orf45*, *DNAH10OS* (Knowles, McLysaght, 2009); 16 новых генов у *Drosophila melanogaster* (Chen et al., 2010); ген *Poldi* у *Mus musculus* (Heinen et al., 2009); ген *MDF1* у *Saccharomyces cerevisiae* (Li et al., 2010). Тем не менее, наибольшей достоверностью считаются исследования, в которых анализ синтенных участков в геномах аутгрупп совмещают с методом геномной филостратиграфии.

ХАРАКТЕРНЫЕ ОСОБЕННОСТИ ГЕНОВ, ВОЗНИКШИХ *DE NOVO*

Возникшие *de novo* гены, обнаруженные у разных видов, обладают некоторыми общими характеристиками. В сравнении с другими, эти гены быстрее эволюционируют (Toll-Riera et al., 2009; Reinhardt et al., 2013), кодируют более короткие ОРС (Zhao et al., 2014), содержат меньше экзонов (Neme, Tautz 2013; Palmieri et al., 2014) и много микросателлитных последовательностей (Guo et al., 2007; Toll-Riera et al., 2009; Palmierit al., 2014). Биоинформатический анализ предпочитаемых кодонов в генах, возникших *de novo,* показал, что в их последовательностях они встречаются чаще чем в последовательностях генов нкРНК и межгенных регионах (Toll-Riera et al., 2009; Palmieri, et al., 2014). Также отмечают следующие общие свойства возникших *de novo* генов: более низкий уровень экспрессии, по сравнению с другими генами, но более высокий, по сравнению с межгенными регионами (Donoghue et al., 2011; Palmieri, et al., 2014; Zhao et al., 2014; Heames et al., 2020), высокую тканеспецифичность (Levine et al., 2006; Toll-Riera et al., 2009; Donoghue et al., 2011; Heames et al., 2020) и частую локализацию на Х-хромосоме (Levine et al., 2006; Palmieri et al., 2014; Zhao et al., 2014). Так, относительно высокий уровень экспрессии генов, возникших *de novo*, по сравнению с другими тканями был обнаружен: в репродуктивной системе самцов плодовой мушки *Drosophila* (Begun et al., 2006, 2007; Zhao et al., 2014); семенниках, селезенке и тимусе домовой мыши (Bekpen et al., 2018); семенниках и коре головного мозга человека (Wu et al., 2011; Pertea et al., 2018). Белковые продукты *de novo*-генов могут восприниматься иммунной системой как чужеродные, что будет приводить к аутоиммунной реакции, поэтому предполагают, что более высокий уровень экспрессии новых генов в мозге и семенниках млекопитающих обеспечивается иммунными привилегиями этих органов, а экспрессия в селезенке и тимусе предшествует экспрессии в других тканях, так как в этих органах происходит адаптация новых белков к иммунному ответу (Bekpen, Tautz, 2018).

В ряде работ показано, что гены, возникшие *de novo*, часто кодируют белки, неспособные формировать выраженную трехмерную структуру, т.е. для этих белков характерна высокая внутренняя неупорядоченность. Некоторые авторы предполагают, что это свойство связано с высоким содержанием нуклеотидов G и C в геномах исследуемых организмов, так как GC-богатые последовательности кодируют более неупорядоченные белки (Basile et al., 2017). Примерами этому являются исследования GC-богатых геномов плодовой мушки *Drosophila*, простейшего *Leishmania major* и дрожжей *Lachancea*, для которых продемонстрирована высокая внутренняя неупорядоченность белков, кодируемых генами, возникшими *de novo* (Bitard-Feildel et al., 2015; Mukherjee et al., 2015; Wilson et al., 2017; Heames et al., 2020), в то время как уровень внутренней неупорядоченности белков, кодируемых генами, возникшими *de novo* в GC-бедном геноме дрожжей *Saccharomyces cerevisiae*, низкий (Carvunis et al., 2012; Ekman, Elofsson, 2010; Basile et al., 2017).

Также одной из гипотез, объясняющих высокий уровень неупорядоченности белков, кодируемых генами, возникшими *de novo*, является гипотеза преадаптации (Ángyán et al., 2012; Wilson et al., 2017; Tretyachenko et al., 2017), согласно которой высокий уровень внутренней неупорядоченности новых белков снижает их способность к агрегации, уменьшая тем самым их потенциальный вред для клетки. Это дает возможность преадаптироваться новым белкам и кодирующих их генам в популяции, до того, как они зафиксируются в ней окончательно. Однако в исследованиях, проведенных на дрожжах *S. cerevisiae* (Vakirlis et al., 2018) и мышах (Schmitz Bornberg-Bauer,

2018) показано, что высокая внутренняя неупорядоченность белков, кодируемых генами, возникшими *de novo*, сопоставима с белками консервативных генов. Тем не менее, в GC-богатых районах реже встречаются стоп-кодоны и сигналы полиаденилирования (AATAAA), что также может способствовать возникновению генов *de novo* (Casola, 2018; Heames et al., 2020). У человека GC-богатые районы более транскрипционно активны (Lercher et al., 2003), а сайты связывания транскрипционных факторов часто обогащены GC-основаниями (Wang et al., 2012). В работе Касола (Casola, 2018) показано, что у домовой мыши только 20 из 152 генов, возникших *de novo*, кодируют белки с высокой внутренней неупорядоченностью за счет оверпринтинга (частичного использования одной из альтернативных ОРС более древнего гена). В работе Доулинга (Dowling et al., 2020) показано, что для человека характерен повышенный, не меняющийся в ходе эволюции приматов, уровень внутренней неупорядоченности белков, кодируемых новыми генами, которые перекрываются с GC-богатыми альтернативными ОРС более древних генов. Следует отметить, что частое формирование перекрытий генов, возникших *de novo*, с другими генами описано во многих работах у приматов и грызунов (Knowles, McLysaght, 2009; Murphy, McLysaght, 2012; Neme, Tautz 2013; Ruiz-Orera et al., 2015; McLysaght, Guerzoni, 2015; Xie et al., 2019), и по оценке Касола может быть почти в шесть раз более частым, чем формирование перекрывающихся пар среди более древних генов (Casola, 2018). Ряд других работ показывает, что белки, кодируемые генами *de novo*, не отличаются от более древних по способности к агрегации, у них не обнаруживают признаки преадаптации, и фиксирование в популяции происходит скорее стохастически, а не под действием отбора (Casola, 2018; Dowling et al., 2020). В работе Нелли-Тибо и Ландри (Nielly-Thibault, Landry, 2019) показано, что высокая внутренняя неупорядоченность новых белков не является следствием действия отбора в сторону снижения их способности к агрегации, это происходит в результате нейтральных процессов возникновения и фиксации в популяции новых генов. Таким образом, вопрос, является ли высокая внутренняя неупорядоченность белков, кодируемых *de novo*-генами, их общей характеристикой, до сих пор остается дискуссионным. Тем не менее внутренняя неупорядоченность новых белков отражает склонность к возникновению кодирующих их генов в GC-богатых транскрипционно-активных участках и вероятно зависит от специфики генома исследуемого организма и алгоритмов поиска новых генов.

Районы генов, возникших *de novo*, имеют характерные модификации хроматина. Анализ генов, возникших *de novo*, у растения *Arabidopsis thaliana* показал, что они часто обладают высоким уровнем метилирования ДНК и почти лишены модификаций гистонов (Li et al., 2016). Авторы также отмечают, что паттерн метилирования генов, возникших *de novo*, стабильно наследуется, а уровень их метилирования может быть опосредован наличием большого числа сайтов, соответствующих малым интерферирующим РНК длиной 24 нуклеотида, играющих ключевую роль в РНК-опосредованном метилировании ДНК. Предполагается, что низкий уровень экспрессии генов, возникших *de novo*, опосредованный высоким уровнем метилирования, позволяет им распространяться и фиксироваться в популяции. Исследования, проведенные на дрожжах, демонстрируют, что большинство генов, возникших *de novo*, локализуется в “горячих” точках рекомбинации, которые обычно лишены нуклеосом (Vakirlis et al., 2018). В исследовании, проведенном на нематоде *Pristionchus pacificus* показано, что точки старта транскрипции эволюционно новых генов часто имеют эпигенетические характеристики энхансеров, а не промотеров, как у консервативных генов (Werner et al., 2018).

РАСПРОСТРАНЕННОСТЬ ГЕНОВ, ВОЗНИКШИХ *DE NOVO*

Оценка числа генов, возникших *de novo*, зависит от нескольких параметров. Она может определяться методом поиска (филостратиграфия, анализ синтенных регионов или более сложные комбинированные методы), зависеть от самого понятия “ген, возникший *de novo*” (например, из анализа могут быть исключены гены, перекрывающиеся с более консервативными генами или имеющие относительно сложную экзон-интронную структуру), доступности секвенированных геномов и транскриптомов видов аутгрупп, включения в анализ экспериментальных подтверждений функций исследуемых генов. Разнообразие условий и методологических подходов приводит к разным оценкам численности генов, возникших *de novo*, даже у одних и тех же организмов. Например, у *Drosophila melanogaster* в разных исследованиях обнаружено 5 (Levine et al., 2006), 2 (Zhou et al., 2008), 16 (Chen et al., 2010), 248 (Zhao et al., 2014), 66 (Heames et al., 2020) генов, возникших *de novo*; у домовой мыши обнаружено 69 (Murphy, McLysaght, 2012), 773 (Neme, Tautz, 2013), 152 (Casola, 2018) таких гена; у человека – 3 (Knowles, McLysaght, 2009) и 66 (Wu et al., 2011). Следует отметить, что значительная часть генов, обнаруженных в одном исследовании, часто не подтверждается в других, так как авторы используют разные по чувствительности методы поиска отдаленных гомологов. Но даже используя наиболее жесткие параметры поиска гомологичных последовательностей большинство авторов еди-

ны во мнении, что, не смотря на высокую степень дивергенции, не позволяющую обнаружить эволюционные гомологи исследуемой последовательности (Vakirlis et al., 2020), процесс возникновения генов *de novo* действительно существует.

ПРЕДПОЛАГАЕМЫЕ МЕХАНИЗМЫ ВОЗНИКНОВЕНИЯ И ЭВОЛЮЦИИ ГЕНОВ *DE NOVO*

Возникновение генов *de novo* в эволюции может происходить в соответствии с несколькими моделями, которые не являются взаимоисключающими.

Транскрипция предшествует возникновению ОРС. Согласно этой модели, начальным этапом появления гена *de novo* является стабильная, но малоинтенсивная транскрипция некодирующего участка генома (Schlötterer, 2015). В пользу этой модели говорят данные исследований, в которых показано, что большая часть генома, не несущая аннотированных генов эукариот способна транскрибироваться, что приводит к возникновению пула длинных некодирующих РНК (нкРНК), значительная часть которых связывается с рибосомами (Kapranov et al., 2007; Wilson, Masel, 2011; Clark et al., 2011; Ruiz-Orera et al., 2014) и вероятно транслируется в виде коротких пептидов (Ruiz-Orera et al., 2014; Ingolia et al., 2014). В процессе эволюции мутации в генах длинных нкРНК могут приводить к удлинению коротких открытых рамок считывания, которые в дальнейшем поддерживаются отбором. Также возможно, что гены нкРНК уже несут относительно длинные рамки считывания, которые прерываются стоп-кодонами. В этом случае мутации, затрагивающие стоп-кодоны могут привести к возникновению полноразмерной рамки считывания (Schlötterer, 2015). Существует ряд исследований, проведенных на разных группах организмов, в которых показано, что возникновению новой ОРС и ее трансляции предшествовала активная транскрипция этого региона (Cai et al., 2008; Carvunis et al., 2012; Reinhardt et al., 2013; Zhang et al., 2019; Schmitz et al., 2020). Особую роль в возникновении генов *de novo* могут играть двунаправленные промоторы и энхансеры (Wu, Sharp 2013; Majic, Payne, 2020). Так, показано, что в районе энхансеров часто транскрибируются так называемые энхансерные РНК (De Santa et al., 2010; Kim et al., 2010; Notani, Rosenfeld, 2016; Haberle, Stark, 2018), которые предположительно могут создавать положительную обратную связь, стимулирующую их собственную транскрипцию, что в конечном итоге приведет к возникновению нового гена (Wu, Sharp, 2013). Так, если эти энхансерные РНК несут даже короткие ОРС, то их относительно высокий уровень экспрессии (возникающий благодаря положительной обратной связи) может способствовать взаимодействиям с рибосомой и возможной трансляции. Но, так как энхансеры активны только в определенных типах клеток (He et al., 2014), трансляция энхансерных РНК будет ограничена только ими, что может способствовать действию отбора в направлении стабилизации экспрессии “полезных” новых белков и возникновению новых генов (Wilson, Masel, 2011). Способность энхансеров в ходе эволюции трансформироваться в промоторы (Carelli et al., 2018) также будет способствовать стабилизации экспрессии нового гена, если он будет обладать адаптивными преимуществами. Таким образом энхансеры могут способствовать как возникновению генов *de novo*, создавая геномное окружение, благоприятствующее транскрипции, так и включению новых генов в существующие генные сети (Majic, Payne, 2020).

ОРС первична по отношению к транскрипции. В данной модели предполагается, что в геноме есть множество неактивных ОРС и для начала их транскрипции и трансляции необходимо возникновение рядом регуляторного элемента (например, промотора) (Schlötterer, 2015). В пользу этой модели говорят результаты исследований, в которых показано, что в геномах эукариот находится множество предсказанных ОРС достаточно длинных для того, чтобы кодировать функционально-значимые пептиды (Carvunis et al., 2012; Zhao et al., 2014). Анализ 248 полиморфно-экспрессирующихся генов, возникших *de novo* у *Drosophila melanogaster*, позволяет предположить, что они возникли из межгенных нетранскрибируемых ОРС (Zhao et al., 2014). Ген тресковых рыб *AFGP*, кодирующий гликопротеин-антифриз, является примером возникновения гена на основе полностью сформировавшейся ОРС, которая стала транскрибироваться после ее предполагаемой транслокации в район, содержащий промотор (Zhuang et al., 2019). Следует отметить, что как минимум у конститутивно экспрессирующихся генов дрожжей большая часть информации о регуляции транскрипцией мРНК закодирована в ОРС, а не в промоторах (Espinar et al., 2018), на основании чего можно предположить, что ОРС генов, возникших *de novo*, могли сочетать в себе информацию о белке и о регуляции собственной транскрипцией, а их промоторы возникали позже. Можно предположить, что энхансеры или переходные от энхансеров к промоторам регуляторные элементы в этом случае тоже играют роль активаторов транскрипции “молчащих” рамок считывания, поскольку энхансеры способны регулировать экспрессию многих дистанционно-удаленных районов генома.

Возникновение генов de novo из прото-генов. Эта модель основана на предположении, что в геноме существуют множество транскрибируемых и транслируемых рамок считывания, лежащих в межгенных участках (Carvunis et al., 2012). Такие ОРС могут стать так называемыми “прото-генами”, способствующими адаптации организма пу-

тем реализации генетической информации, которая обычно скрыта в межгенных участках. Некоторые из прото-генов иногда могут сохраняться в ходе эволюции и становится генами *de novo*, если их экспрессия дает адаптивные преимущества организму. В отличие от псевдогенов, прото-гены не обнаруживают гомологии к известным генам, возникают из негенных последовательностей и обладают промежуточными характеристиками (уровень экспрессии, кодирующий потенциал ОРС, длина, паттерн модификации хроматина, возможность белок-белковых и межгенных взаимодействий и др.) между генами и некодирующими участками. Таким образом формируется континуум между некодирующими участками генома, прото-генами и генами. Модель возникновения генов *de novo* за счет прото-генов хорошо подтверждается результатами исследований на арабидопсисе (Li et al., 2016), дрожжах (Carvunis et al., 2012; Abrusán, 2013) и дрозофиле (Heames et al., 2020).

Гипотеза возникновения генов de novo "Out of testis" ("Из семенника"). Независимо от того, как возникают гены *de novo*, предполагается, что основным регионом их экспрессии у животных являются семенники. В ряде исследований, проведенных на дрозофиле и позвоночных, самый высокий уровень экспрессии генов, возникших *de novo*, отмечается в семенниках (Levine et al., 2006; Begun et al., 2006, 2007; Zhao et al., 2014; Wu et al., 2011; Villanueva-Cañas et al., 2017; Neme, Tautz, 2016). Результаты этих исследований в совокупности с данными о высокой скорости эволюции генов, связанных с размножением (Swanson, Vacquier, 2002; Clark et al., 2006) привели к возникновению гипотезы "Out of testis" ("Из семенника") о решающей роли полового отбора в возникновении генов *de novo*. Было высказано предположение о том, что семенник-специфичные гены, возникшие *de novo*, по неизвестным причинам предпочтительнее сохраняются в ходе эволюции (Palmieri et al., 2014). Предполагается, что высокая транскрипционная активность в семенниках млекопитающих, возникающая за счет высокой экспрессии белков, ее осуществляющих (Schmidt, 1996) и большого количества открытого хроматина (Kleene, 2001), а также иммунная привилегированность семенников, создают необходимые условия для возникновения генов *de novo* (Kaessmann, 2010; Oss, Carvunis, 2019; Zhang, Zhou, 2019).

ПРОБЛЕМА ФУНКЦИОНАЛЬНОСТИ ГЕНОВ, ВОЗНИКШИХ *DE NOVO*

Даже проследив эволюционный путь определенной последовательности, приводящий к возникновению нового гена, одной из проблем при поиске генов, возникших *de novo*, является отсутствие общепринятой точки зрения на то, какой момент времени на филогенетическом древе можно считать моментом возникновения гена. Одной из причин этого является отсутствие единого мнения о том должен ли ген формироваться в полностью некодирующем участке генома (Oss, Carvunis, 2019). Другая проблема, затрудняющая анализ генов, возникших *de novo*, состоит в том, что исследуемая последовательность предполагаемого гена должна соответствовать понятию "ген". Общепринято, что настоящий ген должен кодировать функциональный продукт: белок, или РНК (в случае РНК-генов). Однако, существуют разные взгляды на то, что определяет функцию гена, отчасти в зависимости от того, оценивается ли данная последовательность с позиций генетических, биохимических или эволюционных подходов (Doolittle et al., 2014; Kellis et al., 2014; McLysaght, Hurst, 2016).

Считается, что истинные гены, возникшие *de novo*, должны экспрессировать продукт (на уровне белка или РНК) (Schlötterer, 2015), создавая таким образом предпосылки для воздействия на них естественного отбора. Экспрессия на уровне белка и/или РНК, возникшая *de novo* внутри последовательности ДНК, является важным критерием для придания ей статуса "ген". Экспрессия отдельных генов на уровне РНК может быть подтверждена как давно используемыми стандартными методами (ПЦР с обратной транскрипцией, ПЦР в реальном времени, нозерн-блоттинг, анализ РНК, основанный на ее защите с помощью комплементарной РНК от действия РНКаз), так и более современными широкомасштабными техниками (секвенирование РНК-библиотек). Оценить экспрессию на уровне белка можно, использовав стандартные методики (вестерн-блоттинг, масс-спектрометрия), или широкомасштабные (рибосомный профайлинг) (Ingolia et al., 2009). При этом достоверным экспериментальным подтверждением возникновения гена *de novo* является отсутствие экспрессии синтенного района у видов аутгруппы (Andersson et al., 2015).

Наличие экспрессии в исследуемом районе ДНК, конечно, не является единственным критерием для выяснения функциональности предполагаемого гена, возникшего *de novo*. Более значимыми в этом случае являются результаты генетических экспериментов, таких как нокаут (совокупность методик, позволяющих полностью "выключить" ген) и нокдаун (совокупность методик, позволяющих искусственно снизить экспрессию конкретного гена), так как на основе них можно проследить фенотипические изменения в ответ на нарушение экспрессии исследуемого района (Kellis et al., 2014). Так, для гена *QQS* у *Arabidopsis thaliana* в результате нокдауна с помощью РНК-интерференции было продемонстрировано его участие в метаболизме крахмала (нокдаун *QQS*

приводил к повышенному содержанию крахмала в листьях в конце световой фазы) (Li et al., 2009); в работе Чена и соавт. (Chen et al., 2010) показано, что нокдаун возникших *de novo* генов *Drosophila melanogaster CG9284*, *CG31882* и *CG30395* приводит к летальности. Тем не менее, в случае масштабных биоинформационных скринингов геномов подобные эксперименты обычно не проводятся, но для оценки функциональной значимости предполагаемых генов, возникших *de novo*, в целях получения предварительной информации удобно использовать дополнительные скрининги белок-белковых и межгенных взаимодействий. Для скрининга белок-белковых взаимодействий могут использованы, например, такие методы, как дрожжевой дигибридный скрининг и матрично-активированная лазерная десорбция/ионизация (MALDI-TOF). Оценка генетических взаимодействий исследуемого гена может быть проведена с помощью базы данных BioGRID (Stark et al., 2011). Так, используя данные MALDI-TOF/TOF для белка, кодируемого возникшим *de novo* геном *MDF1* у дрожжей *S. cerevisiae* было предсказано, а в дальнейшем доказано методом коиммунопреципитации, его взаимодействие с белком Snf1p, которое в дальнейшем приводит к ускорению метаболизма глюкозы (Li et al., 2014). В работе Абрусан, также проведенной на дрожжах *S. cerevisiae,* с помощью баз данных BioGRID по белок-белковым, генетическим и регуляторным взаимодействиям (YEASTRACT, Abdulrehman et al., 2011), и в работе по оценке вовлеченности разных групп генов в межгенные взаимодействия (Costanzo et al., 2010), показано, что новые гены слабее интегрированы в регуляторные генные сети, чем более древние (Abrusán, 2013).

Кроме того, для доказательства функциональной значимости исследуемого локуса можно использовать эволюционные подходы, оценивающие действие отбора. В случае таксон-специфичных белок-кодирующих генов одним из признаков действия на них отбора является отношение числа несинонимичных к числу синонимичных замен нуклеотидов (*Ka*/*Ks*, *dN*/*dS*), вычисленное в ходе анализа геномов разных видов данного таксона. Это отношение показывает находится ли исследуемый локус под действием нейтрального (=1), отрицательного (<1) или положительного (>1) отбора. Также в случае видоспецифичных белок-кодирующих генов отношение числа несинонимичных к числу синонимичных замен нуклеотидов *pN*/*pS* можно вычислить на основе данных о полиморфизме в разных линиях или популяциях исследуемых видов. С учетом того, что видоспецифичные гены, возникшие *de novo*, по определению не обладают высокой консервативностью, оценка таких признаков может быть затруднена без анализа большого числа секвенированных геномов линий или популяций. Примером этому могут служить три гена домовой мыши *Mus musculus* (*ENSMUSG00000054057 – Udng1*, *ENSMUSG00000053181 – Udng2*, *ENSMUSG00000078518 – Udng3*), возникшие *de novo*, которые имеют доказанные функции в организме (*Udng1* участвует в регуляции поведения и роста костей, *Udng2* участвует в регуляции поведения, *Udng3* участвует в регуляции времени появления второго помета), но не проявляют признаков действия на них отбора (Xie et al., 2019). Поэтому для выявления действия отбора используют другие характеристики: дивергенцию нуклеотидов в синтенных районах, консервативность границ открытых рамок считывания и их кодирующий потенциал, определяемый на основе частоты встречаемости гексамеров нуклеотидов (Ruiz-Orera et al., 2015).

Несмотря на сложности, связанные с выяснением функции генов, возникших *de novo*, накапливается все больше данных, подтверждающих их роль в жизненно важных и патологических процессах. Например, имеются сведения, что специфические для человека гены вовлечены в онкологические процессы. На модели мышей было показано, что ген *NYCM*, который является уникальным для человека и шимпанзе, регулирует патогенез нейробластом (Chen et al., 2013). Другой специфический для приматов ген *PART1*, синтезирующий длинную некодирующую РНК (lncRNA), в одних работах идентифицирован как опухолевый супрессор, в других – как онкоген (Toll-Riera et al., 2009; Lin et al., 2000; Kang et al., 2018). Несколько других генов, возникших *de novo*, специфичных для человека или приматов, например, *PBOV1* (Samusik et al., 2013), *GR6* (Guerzoni, McLysaght, 2016), *MYEOV* (Papamichos et al., 2015), *ELFN1-AS1* (Kozlov, 2016), и *CLLU1* (Knowles, McLysaght, 2009) также связаны с раком. Специфическая экспрессия многих новых генов в человеческом мозге позволяет высказывать смелые предположения о том, что такие гены могут влиять на когнитивные способности человека (Wu et al., 2011). Одним из таких примеров является уже упомянутый ген *FLJ33706*, который демонстрирует повышенную экспрессию в мозге пациентов с болезнью Альцгеймера (Li et al., 2010). В принципе, экспрессия специфичных для приматов генов в мозге эмбриона человека выше по сравнению с экспрессией подобных генов в мозге мыши. (Zhang et al., 2011). Большинство этих генов, некоторые из которых возникли *de novo*, экспрессируются в неокортексе, который, как известно, отвечает за формирование многих когнитивных способностей человека. Кроме того, они обнаруживают признаки положительного отбора и зачастую участвуют в регуляции транскрипции (Zhang et al., 2011). Помимо своей роли в процессах канцерогенеза, имеются сведения, что возникшие *de novo* гены млекопитающих связаны с работой

иммунной системы (Toll-Riera et al., 2009; Villanueva-Cañas et al., 2017).

ЗАКЛЮЧЕНИЕ

В заключение отметим, что возникшие *de novo* гены важны не только с точки зрения эволюционной биологии – были высказаны предположения, что новые гены, включая те, которые сформировались *de novo*, могут играть важную роль в накоплении специфичных видовых признаков и особенностей (Tautz, Domazet-Lošo, 2011; McLysaght, Guerzoni, 2015; Chen et al., 2013). Сложность заключается в том, что у многих видоспецифических генов отсутствуют аннотированные функции (Villanueva-Cañas et al., 2017). Тем не менее, накопленные данные о преимущественной экспрессии генов *de novo* в семенниках (см. выше), указывающие на их роль в репродукции, вместе с данными о функциях новых генов и их связи с различными заболеваниями у человека и жизненно важными процессами у разных групп организмов говорят в пользу предположения об их роли в накоплении специфичных для каждого вида признаков. Поскольку в настоящее время функции многих, возникших *de novo*, генов человека (и других организмов) остаются не полностью охарактеризованными, необходимы работы, направленные на оценку их конкретного вклада в здоровье и развитие.

ГЛОССАРИЙ

Аутгруппа – филогенетически удаленная от исследуемой группы сестринских таксонов группа, происходящая от общего предка, служит в качестве точки сравнения для исследуемой группы таксонов.

нкРНК – не кодирующая РНК

Ортологи – гомологичные, то есть имеющие общее эволюционное происхождение, схожую структуру, и выполняющие схожую функцию, гены у разных видов организмов.

ОРС – Открытая Рамка Считывания – участок гена, кодирующий белок.

Синтения – нахождение генетических локусов на одной и той же хромосоме, вне зависимости от того, являются ли они сцепленными по данным анализа на сцепленное наследование.

Синтенные блоки – консервативные участки сравниваемых геномов, в которых сохраняется порядок расположения исследуемых элементов.

Скрытые модели Маркова – статистические модели, которые используют для распознавания генов, моделирования их структуры, моделирования семейств последовательностей и др.

BLAST – Basic Local Alignment Search Tool – программное обеспечение, позволяющее находить области сходства между последовательностями белков или нуклеотидов.

CS-BLAST (Context-Specific BLAST) – программное обеспечение расширяющее чувствительность BLAST по поиску схожих аминокислотных последовательностей.

PSI-BLAST (Position-Specific Iterated BLAST) – программное обеспечение, расширяющее возможности BLAST по поиску схожих аминокислотных последовательностей, позволяет пользователю находить сильно филогенетически удаленные гомологичные белки.

k-меры – часть последовательности нуклеотидов определенной длины k.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при финансовой поддержке гранта РФФИ (проект № 20-04-00272а) и в рамках раздела Государственного задания ИБР РАН 2021 года № 0088-2021-0007 “Молекулярно-генетические механизмы регуляции клеточной дифференцировки и морфогенеза”.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая статья не содержит описания выполненных автором исследований с участием людей или использованием животных в качестве объектов.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют, что какой-либо конфликт интересов отсутствует.

ИНФОРМАЦИЯ О ВКЛАДЕ АВТОРОВ

Автор Р.О. Черезов проводил анализ мировой литературы и написание основного текста статьи. Автор Ю.Е. Воронцова участвовала в редактировании и обсуждении текста статьи. О.Б. Симонова инициировала написание обзора и редактировала текст.

СПИСОК ЛИТЕРАТУРЫ

Abdulrehman D., Monteiro P.T., Teixeira M.C. et al. YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface // Nucleic Acids Res. 2011. V. 39. P. D136–D140.

Abrusán G. Integration of new genes into cellular networks, and their structural maturation // Genetics. 2013. V. 195. № 4. P. 1407–1417.

Altschul S.F., Gish W., Miller W. et al. Basic local alignment search tool // J. Mol. Biol. 1990. V. 215. № 3. P. 403–410.

Altschul S.F., Madden T.L., Schäffer A.A. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs // Nucleic Acids Res. 1997. V. 25. № 17. P. 3389–3402.

Andersson D.I., Jerlström-Hultqvist J., Näsvall J. Evolution of new functions *de novo* and from preexisting genes // Cold Spring Harb. Perspect. Biol. 2015. V. 7. № 6. P. a017996.

Ángyán A.F., Perczel A., Gáspári Z. Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? // FEBS Lett. 2012. V. 586. № 16. P. 2468–2472.

Basile W., Sachenkova O., Light S. et al. High GC content causes orphan proteins to be intrinsically disordered // PLoS Comp. Biol. 2017. V. 13. № 3. P. e1005375.

Begun D.J., Lindfors H.A., Kern A.D. et al. Evidence for *de novo* evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* Clade // Genetics. 2007. V. 176. P. 1131–1137.

Begun D.J., Lindfors H.A., Thompson M.E. et al. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags // Genetics. 2006. V. 172. P. 1675–1681.

Bekpen C., Xie C., Tautz D. Dealing with the adaptive immune system during *de novo* evolution of genes from intergenic sequences // BMC Evol. Biol. 2018. V. 18.

Bitard-Feildel T., Heberlein M., Bornberg-Bauer E. et al. Detection of orphan domains in *Drosophila* using “hydrophobic cluster analysis” // Biochimie. 2015. V. 119. P. 244–253.

Cai J., Zhao R., Jiang H. et al. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae* // Genetics. 2008. V. 179. № 1. P. 487–496.

Carelli F.N., Liechti A., Halbert J. et al. Repurposing of promoters and enhancers during mammalian evolution // Nature Comm. 2018. V. 9. № 1. P. 4066.

Carvunis A.-R., Rolland T., Wapinski I. et al. Proto-genes and *de novo* gene birth // Nature. 2012. V. 487. P. 370–374.

Casola C. From *de novo* to “*de nono*”: The majority of novel protein-coding genes identified with phylostratigraphy are old genes or recent duplicates // Genome Biol. Evol. 2018. V. 10. № 11. P. 2906–2918.

Costanzo M., Baryshnikova A., Bellay J. et al. The genetic landscape of a cell // Science. 2010. V. 327. № 5964. P. 425–431.

Chen S., Krinsky B.H., Long M. New genes as drivers of phenotypic evolution // Nat. Rev. Genet. 2013. V. 14. № 9. P. 645–660.

Chen S., Zhang Y.E., Long M. New genes in *Drosophila* quickly become essential // Science. 2010. V. 330. P. 1682–1685.

Clark M.B., Amaral P.P., Schlesinger F.J. et al. The reality of pervasive transcription // PLoS Biol. 2011. V. 9. № 7. P. e1000625.

Clark N.L., Aagaard J.E., Swanson W.J. Evolution of reproductive proteins from animals and plants // Reproduction. 2006. V. 131. № 1. P. 11–22.

De Santa F., Barozzi I., Mietton F. et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers // PLoS Biology. 2010. V. 8. № 5. P. e1000384.

Domazet-Loso T., Brajković J., Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages // Trends Genet. 2007. V. 23. № 11. P. 533–539.

Domazet-Lošo T., Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns // Nature. 2010. V. 468. № 7325. P. 815–818.

Donoghue M.T., Keshavaiah C., Swamidatta S.H. et al. Evolutionary origins of *Brassicaceae* specific genes in *Arabidopsis thaliana* // BMC Evol. Biol. 2011. V. 11. № 1. P. 47.

Doolittle W.F., Brunet T.D.P., Linquist S. et al. Distinguishing between “function” and “effect” in genome biology // Genome Biol. Evol. 2014. V. 6. № 5. P. 1234–1237.

Dowling D., Schmitz J.F., Bornberg-Bauer E. Stochastic gain and loss of novel transcribed open reading frames in the human lineage // Genome Biol. Evol. 2020. V. 12. № 11. P. 2183–2195.

Dujon B. The yeast genome project: what did we learn? // Trends Genet. 1996. V. 12. № 7. P. 263–270.

Ekman D., Elofsson A. Identifying and quantifying orphan protein sequences in fungi // J. Mol. Biol. 2010. V. 396. № 2. P. 396–405.

Elhaik E., Sabath N., Graur D. 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. 2006. V. 23. № 1. P. 1–3.

Espinar L., Tamarit M.À.S., Domingo J., Carey L.B. Promoter architecture determines cotranslational regulation of mRNA // Genome Res. 2018. V. 28. № 4. P. 509–518.

Gehrmann T., Reinders M.J.T. Proteny: discovering and visualizing statistically significant syntenic clusters at the proteome level // Bioinformatics. 2015. V. 31. № 21. P. 3437–3444.

Ghiurcuta C.G., Moret B.M.E. Evaluating synteny for improved comparative studies // Bioinformatics. 2014. V. 30. № 12. P. i9–i18.

Grasse P.-P. Evolution of Living Organisms. London: Academic Press, 1977. 308 p.

Guerzoni D., McLysaght A. De novo genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting // Genome Biol. Evol. 2016. V. 8. № 4. P. 1222–1232.

Guo W.-J., Li P., Ling J. et al. Significant comparative characteristics between orphan and nonorphan genes in the rice (*Oryza sativa* L.) genome // Comp. Func. Genomics. 2007. V. 2007. P. 21676.

Haberle V., Stark A. Eukaryotic core promoters and the functional basis of transcription initiation // Nat. Rev. Mol. 2018. V. 19. № 10. P. 621–637.

Haldane J.B.S. The part played by recurrent mutation in evolution // Am. Nat. 1933. V. 67. № 708. P. 5–19.

He B., Chen C., Teng L. et al. Global view of enhancer-promoter interactome in human cells // Proc. Natl. Acad. Sci. USA. 2014. V. 111. № 21. P. E2191–E2199.

Heames B., Schmitz J., Bornberg-Bauer E. A Continuum of evolving *de novo* genes drives protein-coding novelty in *Drosophila* // J. Mol. Evol. 2020. V. 88. № 4. P. 382–398.

Heinen T.J.A.J., Staubach F., Häming D. et al. Emergence of a new gene from an intergenic region // Curr. Biol. 2009. V. 19. № 18. P. 1527–1531.

Husnik F., McCutcheon J.P. Functional horizontal gene transfer from bacteria to eukaryotes // Nat. Rev. Microbiol. 2018. V. 16. № 2. P. 67–79.

Ingolia N.T., Brar G.A., Stern-Ginossar N. et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes // Cell Rep. 2014. V. 8. № 5. P. 1365–1379.

Ingolia N.T., Ghaemmaghami S., Newman J.R.S. et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling // Science. 2009. V. 324. № 5924. P. 218–223.

Jacob F. Evolution and tinkering // Science. 1977. V. 196. P. 1161–1166.

Kaessmann H. Origins, evolution, and phenotypic impact of new genes // Genome Res. 2010. V. 20. № 10. P. 1313–1326.

Kang M., Ren M., Li Y., Fu Y. et al. Exosome-mediated transfer of lncRNA PART1 induces gefitinib resistance in esophageal squamous cell carcinoma via functioning as a competing endogenous RNA // J. Exp. Clin. Cancer Res. 2018. V. 37.

Kapranov P., Willingham A.T., Gingeras T.R. Genome-wide transcription and the implications for genomic organization // Nat. Rev. Genet. 2007. V. 8. № 6. P. 413–423.

Kellis M., Wold B., Snyder M.P. et al. Defining functional DNA elements in the human genome // Proc. Natl. Acad. Sci. USA. 2014. V. 111. № 17. P. 6131–6138.

Kim T.-K., Hemberg M., Gray J.M. et al. Widespread transcription at neuronal activity-regulated enhancers // Nature. 2010. V. 465. P. 182–187.

Knowles D.G., McLysaght A. Recent *de novo* origin of human protein-coding genes // Genome Res. 2009. V. 19. № 10. P. 1752–1759.

Kozlov A.P. Expression of evolutionarily novel genes in tumors // Infect. Agents Cancer. 2016. V. 11. № 34.

Kröger H., Donner I., Skiello G. Influence of a new virostatic compound on the induction of enzymes in rat liver // Arzneimittelforschung. 1975. V. 25. № 9. P. 1426–1429.

Lercher M.J., Urrutia A.O., Pavlíček A. et al. A unification of mosaic structures in the human genome // Hum. Mol. Genet. 2003. V. 12. № 19. P. 2411–2415.

Levine M.T., Jones C.D., Kern A.D. et al. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression // Proc. Natl. Acad. Sci. USA. 2006. V. 103. № 26. P. 9935–9939.

Li C.-Y., Zhang Y., Wang Z. et al. A human-specific *de novo* protein-coding gene associated with human brain functions // PLoS Comput. Biol. 2010. V. 6. № 3. P. e1000734.

Li D., Dong Y., Jiang Y. et al. A *de novo* originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand // Cell Res. 2010. V. 20. № 4. P. 408–420.

Li D., Yan Z., Lu L. et al. Pleiotropy of the *de novo*-originated gene MDF1 // Sci. Rep. 2014. V. 4. № 1. P. 7280.

Li L., Foster C.M., Gan Q. et al. Identification of the novel protein QQS as a component of the starch metabolic network in *Arabidopsis* leaves // Plant J. 2009. V. 58. № 3. P. 485–498.

Li W., Notani D., Rosenfeld M.G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives // Nat. Rev. Genet. 2016. V. 17. № 4. P. 207–223.

Li Z.-W., Chen X., Wu Q. et al. On the origin of *de novo* genes in *Arabidopsis thaliana* populations // Genome Biol. Evol. 2016. V. 8. № 7. P. 2190–2202.

Lin B., White J.T., Ferguson C. et al. PART-1: a novel human prostate-specific, androgen-regulated gene that maps to chromosome 5q12 // Cancer Res. 2000. V. 60(4). P. 858–863.

Liu D., Hunt M., Tsai I.J. Inferring synteny between genome assemblies: a systematic evaluation // BMC Bioinformatics. 2018. V. 19. № 1. P. 26.

Luis Villanueva-Cañas J., Ruiz-Orera J., Agea M.I. et al. New genes and functional innovation in Mammals // Genome Biol. Evol. 2017. V. 9. № 7. P. 1886–1900.

Majic P., Payne J.L. Enhancers facilitate the birth of *de novo* genes and gene integration into regulatory networks // Mol. Biol. Evol. 2020. V. 37. № 4. P. 1165–1178.

McLysaght A., Guerzoni D. New genes from non-coding sequence: the role of *de novo* protein-coding genes in eukaryotic evolutionary innovation // Philos. Trans. R. Soc. Lond. B Biol. Sci. 2015. V. 370. № 1678. P. 20140332.

McLysaght A., Hurst L.D. Open questions in the study of *de novo* genes: what, how and why // Nat. Rev. Genet. 2016. V. 17. № 9. P. 567–578.

Moyers B.A., Zhang J. Phylostratigraphic bias creates spurious patterns of genome evolution // Mol. Biol. Evol. 2015. V. 32. № 1. P. 258–267.

Moyers B.A., Zhang J. Evaluating phylostratigraphic evidence for widespread *de novo* gene birth in genome evolution // Mol. Biol. Evol. 2016. V. 33. № 5. P. 1245–1256.

Mukherjee S., Panda A., Ghosh T.C. Elucidating evolutionary features and functional implications of orphan genes in *Leishmania major* // Infect. Genet. Evol. 2015. V. 32. P. 330–337.

Muller H.J. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere // Genetica. 1935. V. 17. № 3. P. 237–252.

Murphy D.N., McLysaght A. De novo origin of protein-coding genes in murine rodents // PLoS One. 2012. V. 7. № 11. P. e48650.

Neme R., Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution // BMC Genomics. 2013. V. 14. № 1. P. 117.

Neme R., Tautz D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence // eLife. 2016. V. 5. P. e09977.

Nielly-Thibault L., Landry C.R. Differences between the raw material and the products of *de novo* gene birth can result from mutational biases // Genetics. 2019. V. 212. № 4. P. 1353–1366.

Nishida H. Detection and characterization of fungal-specific proteins in *Saccharomyces cerevisiae* // Biosci. Biotechnol. Biochem. 2006. V. 70. № 11. P. 2646–2652.

Ohno S. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding

sequence // Proc. Natl. Acad. Sci. USA. 1984. V. 81. № 8. P. 2421–2425.

Ohno S., Wolf U., Atkin N.B. Evolution from fish to mammals by gene duplication // Hereditas. 1968. V. 59. № 1. P. 169–187.

Ohno S. Evolution by Gene Duplication. Berlin: Springer-Verlag, 1970. 160 p.

Oss S.B.V., Carvunis A.-R. De novo gene birth // PLoS Genet. 2019. V. 15. № 5. P. e1008160.

Palmieri N., Kosiol C., Schlötterer C. The life cycle of *Drosophila* orphan genes // eLife. 2014. V. 3. P. e01311.

Papamichos S.I., Margaritis D., Kotsianidis I. Adaptive evolution coupled with retrotransposon exaptation allowed for the generation of a human-protein-specific coding gene that promotes cancer cell proliferation and metastasis in both haematological malignancies and solid tumours: the extraordinary case of MYEOV gene // Sci. 2015. V. 2015. P. 984706.

Pertea M., Shumate A., Pertea G. et al. Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise // bioRxiv. 2018. P. 332825.

Potter S.C., Luciani A., Eddy S.R. et al. HMMER web server: 2018 update // Nucleic Acids Res. 2018. V. 46. P. W200–W204.

Rancurel C., Khosravi M., Dunker A.K. et al. Overlapping genes produce proteins with unusual sequence properties and offer insight into *de novo* protein creation // J. Virol. 2009. V. 83. P. 10719–10736.

Ranz J.M., Casals F., Ruiz A. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila* // Genome Res. 2001. V. 11. P. 230–239.

Reinhardt J.A., Wanjiru B.M., Brant A.T. et al. De novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences // PLoS Genet. 2013. V. 9. P. e1003860.

Ruiz-Orera J., Hernandez-Rodriguez J., Chiva C. et al. Origins of *de novo* genes in human and chimpanzee // PLoS Genet. 2015. V. 11. P. e1005721.

Ruiz-Orera J., Messeguer X., Subirana J.A. et al. Long non-coding RNAs as a source of new peptides // eLife. 2014. V. 3. P. e03523.

Samusik N., Krukovskaya L., Meln I., Shilov E., Kozlov A.P. PBOV1 is a human de novo gene with tumor-specific expression that is associated with a positive clinical outcome of cancer // PLoS One. 2013. V. 8. № 2. P. e56162.

Schlötterer C. Genes from scratch – the evolutionary fate of *de novo* genes // Trends Genet. 2015. V. 31. № 4. P. 215–219.

Schmidt E.E. Transcriptional promiscuity in testes // Curr. Biol. 1996. V. 6(7). P. 768–769.

Schmitz J.F., Chain F.J.J., Bornberg-Bauer E. Evolution of novel genes in three-spanned stickleback populations // Heredity. 2020. V. 125. P. 50–59.

Schmitz J.F., Ullrich K.K., Bornberg-Bauer E. Incipient *de novo* genes can evolve from frozen accidents that escaped rapid transcript turnover // Nat. Ecol. Evol. 2018. V. 2. № 10. P. 1626–1632.

Stark C., Breitkreutz B.-J., Chatr-aryamontri A. et al. The BioGRID Interaction Database: 2011 update // Nucleic Acids Res. 2011. V. 39. P. D698–D704.

Swanson W.J., Vacquier V.D. The rapid evolution of reproductive proteins // Nat. Rev. Genet. 2002. V. 3. № 2. P. 137–144.

Tautz D., Domazet-Lošo T. The evolutionary origin of orphan genes // Nat. Rev. Genet. 2011. V. 12. № 10. P. 692–702.

Toll-Riera M., Bosch N., Bellora N. et al. Origin of primate orphan genes: a comparative genomics approach // Mol. Biol. Evol. 2009. V. 26. № 3. P. 603–612.

Tretyachenko V., Vymětal J., Bednárová L. et al. Random protein sequences can form defined secondary structures and are well-tolerated *in vivo* // Sci. Rep. 2017. V. 7. № 1. P. 15449.

Vakirlis N., Carvunis A.-R., McLysaght A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes // eLife. 2020. V. 9. P. e53500.

Vakirlis N., Hebert A.S., Opulente D.A. et al. Molecular portrait of *de novo* genes in yeasts // Mol. Biol. Evol. 2018. V. 35. № 3. P. 631–645.

Wang J., Zhuang J., Iyer S. et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors // Genome Res. 2012. V. 22. № 9. P. 1798–1812.

Weisman C.M., Murray A.W., Eddy S.R. Many, but not all, lineage-specific genes can be explained by homology detection failure // PLoS Biol. 2020. V. 18. № 11. P. e3000862.

Werner M.S., Sieriebriennikov B., Prabh N. et al. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation // Genome Res. 2018. V. 28. № 11. P. 1675–1687.

Wilson B.A., Foy S.G., Neme R. et al. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth // Nat. Ecol. Evol. 2017. V. 1. № 6. P. 0146.

Wilson B.A., Masel J. Putatively noncoding transcripts show extensive association with ribosomes // Genome Biol. Evol. 2011. V. 3. P. 1245–1252.

Wu D.-D., Irwin D.M., Zhang Y.-P. De novo origin of human protein-coding genes // PLoS Genet. 2011. V. 7. № 11. P. e1002379.

Wu X., Sharp P.A. Divergent transcription: a driving force for new gene origination? // Cell. 2013. V. 155. № 5. P. 990–996.

Xie C., Bekpen C., Künzel S. et al. Studying the dawn of *de novo* gene emergence in mice reveals fast integration of new genes into functional networks // bioRxiv. 2019. P. 510214.

Zhang J.-Y., Zhou Q. On the regulatory evolution of new genes throughout their life history // Mol. Biol. Evol. 2019. V. 36. № 1. P. 15–27.

Zhang L., Ren Y., Yang T. et al. Rapid evolution of protein diversity by *de novo* origination in *Oryza* // Nat. Ecol. 2019. V. 3. № 4. P. 679–690.

Zhang W., Gao Y., Long M., Shen B. Origination and evolution of orphan genes and *de novo* genes in the genome of

Caenorhabditis elegans // Sci. China Life Sci. 2019. V. 62. P. 579–593.

Zhang Y.E., Landback P., Vibranovski M.D., Long M. Accelerated recruitment of new brain development genes into the human genome // PLoS Biol. 2011. V. 9. № 10. P. e1001179.

Zhao L., Saelao P., Jones C.D. et al. Origin and spread of *de novo* genes in *Drosophila melanogaster* populations // Science. 2014. V. 343. № 6172. P. 769–772.

Zhou Q., Zhang G., Zhang Y. et al. On the origin of new genes in *Drosophila* // Gen. Res. 2008. V. 18. № 9. P. 1446–1455.

Zhuang X., Yang C., Murphy K.R. et al. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids // Proc. Natl. Acad. Sci. USA. 2019. V. 116. № 10. P. 4400–4405.

The Phenomenon of *De Novo* Gene Birth and Evolution

R. O. Cherezov[1], Ju. E. Vorontsova[1], and O. B. Simonova[1, *]

[1] *Koltzov Institute of Developmental Biology of Russian Academy of Sciences, ul. Vavilova 26, Moscow, 119334 Russia*
**e-mail: osimonova@hotmail.com*

Evolutionary biologists have always been interested in the origin and evolution of new genes. The most obvious mechanisms of their formation are various kinds of chromosomal and intergenic rearrangements, implying the use of already existing genes as a starting material. The possibility of *de novo* origin of a functional gene within noncoding DNA was not fully rejected, but until recently, it was practically going to zero. Nevertheless, in 1996, after analyzing the genome of the yeast *Saccharomyces cerevisiae*, the first experimental evidence was obtained for the possibility of *de novo* gene birth. Ten years later, genes without homologues, presumably arose *de novo*, were found in *Drosophila*. The relatively high probability of genes arisen *de novo*, assessed in bioinformatics studies, has raised interest in this topic and made the search for them relevant. Now the number of works devoted to the problem of *de novo* gene birth in different organisms, including humans, is constantly growing, demystifying this phenomenon. Nevertheless, many questions still require theoretical and practical research. This review is devoted to problems of finding and characterizing genes that have arisen *de novo* and the proposed mechanisms of their birth.

Keywords: gene, evolution, *de novo*