doi 10.18699/bgrs2026-37

Пайплайн по идентификации доместицированных мобильных элементов

Гусева Ю.[1, 2]*, Резвых А.[1], Фуников С.[1]
[1] *Институт молекулярной биологии им. В.А. Энгельгардта РАН, Москва, Россия*
[2] *Московский физико-технический институт (национальный исследовательский университет), Москва, Россия*
* *guseva.iuo@phystech.edu*

Ключевые слова: мобильные элементы; доместикация; *Tc1/Mariner*; эволюция; сравнительная геномика; *CG14478*; *CG4570*

Мотивация и цель: Появление новых генов способствуют эволюционным инновациям во всех организмах [1]. Одним из важнейших механизмов возникновения новых генов является доместикация или молекулярное «одомашнивание» транспозонов, то есть переход последовательности мобильного элемента (МЭ) из «геномного паразита» в функциональный ген хозяина [1, 2]. Такие события приводят к формированию новых регуляторных и белковых последовательностей в геноме, поэтому их поиск важен для понимания эволюции генома. На данный момент, универсальная база генов, однозначно классифицированных как продукты доместикации отсутствует, а большинство исследований опирались только на гомологию между генами и белками транспозонов. В данной работе описана разработка универсального биоинформатического пайплайна для идентификации генов-кандидатов, возникших в результате доместикации мобильных элементов, основанного на анализе геномов подсемейства *Drosophilinae.* С использованием пайплайна было проведёно детальное исследование генов, гомологичных транспозазам семейства *Tc1/Mariner*, с целью уточнения их функций, степени консервативности, характера отбора и вероятного времени доместикации.
Методы и алгоритмы: Первый этап алгоритма заключался в поиске гомологий между белками мобильных элементов и белок-кодирующими последовательностями (CDS) в геномах исследуемых организмов. Для этого использовался инструмент tblastn с запросом, состоящим из белков мобильных элементов из базы RepatPeps, содержащей в себя около 40 тыс. белковых последовательностей транспозонов различных организмов. Выравнивание выполнялось на CDS-последовательности генов из 43 геномов подсемейства *Drosophilinae*. Для каждого выравнивания оценивалась статистика bitscore, а в дальнейший анализ включались только те выравнивания, у которых значение этой статистики было выше третьего квартиля на 1,5 межквартильного размаха, что позволило существенно снизить вероятность случайных совпадений. На следующем этапе была выполнена дополнительная фильтрация базы RepeatPeps, поскольку в ней присутствовали последовательности с нетипичной для заявленного семейства доменной архитектурой. Идентификация доменных структур генов-кандидатов и белков МЭ производилась с использованием базы доменов CDD и инструмента rpsblast. Белки транспозонов были сгруппированы по соответствующим им семействам, после чего для каждого семейства была рассчитана частота встречаемости каждого идентифицированного домена. Чтобы определить, является ли доменная архитектура конкретного МЭ типичной для его суперсемейства, использовали статистический тест, реализованный в функции xmulti языка программирования R в пакете XNomial. Сравнение набора значений частот доменов у МЭ и распределение частот доменов в его семействе позволило исключить нетипичные для суперсемейства ТЕ-последовательности и оставить только канонические представители семейства. Далее выполнялась проверка генов-кандидатов на соответствие критериям, отличающим гены от ДНК-транспозонов. Критериями для отбора были выбраны: отсутствие терминальных инвертированных повторов (TIR) и таргетных сайтов дупликации (TSD) в генах и фланкирующих их областей; низкая копийность в геноме, исключающая активную экспансию; действие очищающего отбора; наличие экспрессии в соматических клетках; консервативность микросентении и отдельное монофилетическое положение ортологов гена относительно предкового семейства транспозонов. Для валидации работы алгоритма гены, прошедшие отбор, подверглись эволюционно-функциональному анализу. Были определены ортологи генов-кандидатов, гомологичных транспозазам суперсемейства *Tc1/Mariner*, с помощью базы данных OrthoDB, далее на основе их последовательностей было выполнено множественное выравнивание с использованием программы MAFFT, которое использовалось для построения филогенетического дерева и оценки направление действующего на ген отбора по значению dN/dS. Функциональный анализ включал изучение тканеспецифической и стадийной экспрессии по данным modENCODE, построение сетей коэкспрессии и GO-обогащения.
Результаты: Разработанный пайплайн был применён к 43 геномам подсемейства *Drosophilinae* и в результате его работы были отобраны 116 генов, белки которых высоко гомологичны к белкам МЭ. Минимальное значение bitscore составило 37.7 для данных пар гомологий. Нами были разработаны 6 критериев доместикации, которые могли подтвердить, что найденный ген является функциональной частью генома, а не активным МЭ. Так как в предыдущих работах не производилось мультигеномного анализа доместикаций МЭ семейства *Tc1/Mariner*, критерии отбора были применены к группе генов, сходных с транспозазами именно этого суперсемейства транспозонов. В итоге было подтверждено наличие пяти генов, обладающих наиболее значимыми признаками доместикации. Четыре из них (*CG14478*, *CG4570*, *cag* и *toy*) имеют признаки древней доместикации, включая очищающий отбор, стабильную экспрессию, отдельное положение на филогенетическом дереве и

консервативную микросинтенцию. Пятый, ген, ограниченный определенной кладой (*LOC26532114*) в группе видов *Drosophila obscura*, демонстрирует признаки недавнего одомашнивания, характеризующиеся умеренным очищающим отбором и частичным сохранением доменной структуры. Важно отметить, что проведенный анализ показал существование двух возможных молекулярных путей доместикации транспозонов суперсемейства *Tc1/Mariner*: ген *CG14478* сохранил каталитический эндонуклеазный DDE-домен, а ген *CG4570* – HTH-мотив ДНК-связывающего домена. При этом предсказанная трёхмерная структура белка гена *CG1448* высокогомологична структуре белков активных транспозонов и положение DDE-аминокислот, критическое для каталитической активности, сохраняется. По профилям экспрессии и сетям коэкспрессии можно предположить, что *CG14478* связан с регуляцией РНК и метаболизмом хроматина, *CG4570* – с клеточным циклом и организацией хромосом, а для *cag* и *toy* возможны функции, связанные с цилиогенезом и развитием нейронов. Таким образом, разработанный пайплайн позволяет не только находить доместицированные гены, но и предполагать механизм их функционального использования.

Выводы: Нами был разработан и протестирован универсальный воспроизводимый пайплайн для поиска генов-кандидатов в «доместицированные транспозоны». В отличие от подходов, основанных только на сходстве последовательностей, предложенная схема сочетает статистическую фильтрацию, доменный анализ, эволюционные критерии, микросинтению и функциональные данные, что заметно повышает точность выводов об эволюции данных генов. Также было показано, что в геномах исследуемой группы присутствует не менее пяти генов, произошедших от транспозонов семейства *Tc1/Mariner*, которые различаются по времени доместикации и функциональной специализации. Предложенный пайплайн в дальнейшем можно использовать для масштабного поиска событий молекулярной доместикации в других эукариотических геномах и создания открытой базы данных таких генов. Также планируется доработка и оптимизация алгоритмов пайплайна для эффективного выполнение поиска доместицированных мобильных элементов на больших наборах геномов из NCBI. На данный момент, идёт отлаживание процессов наработки белка *CG14478* для определения его каталитической активности in vitro, и, в дальнейшем, исследования функций *CG14478*.

Финансирование: Исследование поддержано РНФ № 22-74-10050-П.

A pipeline for identifying domesticated transposable elements

Guseva I.[1,2]*, Rezvykh A.[1], Funikov S.[1]
[1] *Engelhardt Institute of Molecular Biology of Russian Academy of Sciences, Moscow, Russia*
[2] *Moscow Institute of Physics and Technology (National Research University), Moscow, Russia*
* *guseva.iuo@phystech.edu*

Key words: transposable elements; domestication; *Tc1/Mariner*; evolution; comparative genomics; *CG14478*; *CG4570*

Motivation and Aim: The emergence of new genes contributes to evolutionary innovations in all organisms [1]. One of the most important mechanisms for the emergence of new genes is a molecular “domestication” of transposons, i.e., the transition of a transposable element (TE) sequence from a “genomic parasite” to a functional host gene [1, 2]. Such events lead to the formation of new regulatory and protein sequences in the genome, so their identification is important for understanding genome evolution. Currently, there is no universal database of genes unambiguously classified as a products of domestication, and most studies have relied solely on homology between transposon genes and proteins. This study describes the development of a universal bioinformatical pipeline for identifying candidate genes arising from the domestication of TEs, based on an analysis of the genomes of the *Drosophilinae* subfamily. Using the pipeline, a detailed study of genes homologous to *Tc1/Mariner* family transposases was conducted to clarify their functions, degree of conservation, nature of selection, and probable time of domestication.

Methods and Algorithms: The first stage of the algorithm involved searching for homologies between TE proteins and protein-coding sequences (CDS) in the genomes of the studied organisms. To aim this purpose, the tblastn tool was used with a query consisting of TE proteins from the RepatPeps database, which contains approximately 40,000 transposon protein sequences from various organisms. Alignment was performed on the CDS sequences of genes from 43 genomes of the Drosophilinae subfamily. For each alignment, the bitscore statistic was evaluated, and only those alignments for which this statistic was 1.5 interquartile ranges above the third quartile were included in further analysis, which significantly reduced the probability of random matches. In the next step, additional filtering of the RepeatPeps database was performed, as it contained sequences with domain architectures atypical for the claimed family. The identification of domain structures in candidate genes and TE proteins was performed using the CDD domain database and the rpsblast tool. Transposon proteins were grouped into their corresponding families, the frequency of occurrence of each identified domain was calculated. To determine whether the domain architecture of a specific TE is typical for its superfamily, we used a statistical test implemented in the xmulti function of the R programming language in the XNomial package. Comparing the set of domain frequency values for the TE with the distribution of domain frequencies within its family allowed us to exclude TE sequences atypical for the superfamily and retain only canonical representatives of the family. Next, candidate genes were checked for compliance with criteria distinguishing genes from DNA transposons. The following criteria were selected for selection: the absence of terminal inverted repeats (TIRs) and target sites of duplication (TSDs) in the genes and their flanking regions; low

copy number in the genome, ruling out active expansion; the action of purifying selection; the presence of expression in somatic cells; conservatism of the microsatellite and the separate monophyletic position of the gene's orthologs relative to the ancestral transposon family. To validate the work of the algorithm, the genes that passed the selection underwent evolutionary and functional analyses. Orthologs of candidate genes homologous to transposons of the *Tc1/Mariner* superfamily were identified using the OrthoDB database, subsequently, multiple alignment was performed based on their sequences using the MAFFT program, which was used to construct a phylogenetic tree and assess the direction of selection acting on the gene based on the dN/dS ratio. Functional analysis included the study of tissue-specific and developmental stage-specific expression based on modENCODE data, the construction of co-expression networks, and GO enrichment analysis.

Results: The developed pipeline was applied to 43 genomes of the *Drosophilinae* subfamily, resulting in the identification of 116 genes whose proteins are highly homologous to TE proteins. The minimum bitscore value was 37.7 for these homology pairs. We developed six domestication criteria that could confirm that the identified gene is a functional part of the genome rather than a TE. Since previous studies did not perform a multigene analysis of *Tc1/Mariner* family TE domestications, the selection criteria were applied to a group of genes similar to transposases of this specific transposon superfamily. As a result, the presence of five genes exhibiting the most significant signs of domestication was confirmed. Four of them (*CG14478*, *CG4570*, *cag*, and *toy*) exhibit signs of ancient domestication, including purifying selection, stable expression, a distinct position on the phylogenetic tree, and conservative microsynteny. The fifth gene, restricted to a specific clade (*LOC26532114*) within the *Drosophila obscura* species group, exhibits signs of recent domestication, characterized by moderate purifying selection and partial preservation of the domain structure. It is important to note that the analysis revealed the existence of two possible molecular pathways for the domestication of *Tc1/Mariner* family transposons: the *CG14478* gene retained the catalytic endonuclease DDE domain, while the *CG4570* gene retained the DNA-binding HTH domain. Along these lines, the predicted three-dimensional structure of the *CG1448* protein is highly homologous to the structure of active transposon proteins, and the position of the DDE amino acids, which is critical for catalytic activity, is preserved. Based on expression profiles and co-expression networks, it can be suggested that *CG14478* is associated with RNA regulation and chromatin metabolism, *CG4570* with the cell cycle and chromosome organization, and for *cag* and *toy*, functions related to cytogenesis and neuronal development are possible. Thus, the developed pipeline allows not only for the identification of domesticated genes but also for the prediction of the mechanism of their functional utilization.

Conclusion: We developed and tested a universal, reproducible pipeline for identifying candidate genes as "domesticated transposons". Unlike approaches based solely on sequence similarity, the proposed pipeline combines statistical filtering, domain analysis, evolutionary criteria, microsynteny, and functional data, which significantly improves the accuracy of conclusions regarding the evolution of these genes. It was also shown that the genomes of the studied group contain at least five genes derived from *Tc1/Mariner* family transposons, which differ in terms of domestication time and functional specialization. The proposed pipeline can be used in the future for large-scale screening for molecular domestication events in other eukaryotic genomes and for the creation of an open database of such genes. We also plan to refine and optimize the pipeline's algorithms to efficiently search for domesticated TEs in large sets of genomes from NCBI. Currently, we are fine-tuning the processes for producing the *CG14478* protein to determine its DNA-lytic activity in vitro, and, in the future, to study the functions of *CG14478*.

Funding: The study is supported by the Russian Science Foundation (RSF) grant No. 22-74-10050-P.

Список литературы/References

1. Xia S, Chen J, Arsala D, Emerson JJ, Long M. Functional innovation through new genes as a general evolutionary process. *Nat Genet*. 2025;57(2):295-309. doi:10.1038/s41588-024-02059-0
2. Modzelewski AJ, Chong JG, Wang T, He L. Mammalian genome innovation through transposon domestication. Nature Cell Biology. 2022;24(9):1332-1340. doi:10.1038/s41556-022-00970-4