

Предсказание и анализ длинных некодирующих РНК на основе пантранскриптома

Пронозин А.^{1, 2*}, Афонников Д.^{1, 2}

¹ Институт цитологии и генетики СО РАН, Новосибирск, Россия

² Курчатowski геномный центр ИЦиГ СО РАН, Новосибирск, Россия

* pronozinartem95@gmail.com

Ключевые слова: днРНК; пантранскриптом; кукуруза

Мотивация и цели: Длинные некодирующие РНК (днРНК) представляют собой класс линейных или кольцевых молекул РНК длиной от 200 нуклеотидов, не кодирующих белок. К настоящему времени идентифицировано более полу-миллиона последовательностей днРНК для различных организмов.

Однако, несмотря на растущее число исследований, структурно-функциональные характеристики днРНК по-прежнему остаются малоизученными. Это связано со множеством факторов, которые нужно учитывать при идентификации днРНК. Размер днРНК схож с размером некоторых генов, кодирующих белок, что приводит к ошибкам перепредсказания. Также исследования показывают, что последовательности днРНК претерпевают быструю эволюцию, закономерности которой пока не исследованы. Таким образом, возникает необходимость анализа генного состава сразу у нескольких представителей вида. Для этих целей были предложены концепции пангенома и пантранскриптома.

Однако на сегодняшний день работы, посвященные исследованию пантранскриптомов, в основном направлены на выявление и исследование новых генов, кодирующих белок. Тогда как работ, посвященных исследованию днРНК в масштабах пантранскриптомов, не так много, в особенности для растений.

Данная работа направлена на расширение знаний о пантранскриптоме кукурузы за счет дополнительного анализа днРНК.

Методы и алгоритмы: В работе использовалось 503 библиотеки инбредных линий кукурузы, полученных в работе [1]. Для сборки и анализа пантранскриптома был разработан биоинформатический конвейер, который включает три этапа обработки: контроль качества и сборка транскриптома, разделение транскриптомов на кодирующую и некодирующую части и анализ пантранскриптомов.

Результаты: В результате получено два пантранскриптома кукурузы. Пантранскриптом, содержащий кодирующую часть, насчитывает 33023 гена, кодирующих белок. Пантранскриптом, содержащий некодирующую часть, насчитывает 2215493 днРНК. Для обоих пантранскриптомов была проведена аннотация, определены группы ортологов, охарактеризованы консервативные и варьируемые части. Также отдельно для пантранскриптома с некодирующей частью был проведен поиск новых днРНК путем выравнивания полученных транскриптов с базами данных известных днРНК и анализа экспрессии выявленных новых днРНК. Для пантранскриптома с кодирующей частью был проведен анализ ГО-терминов и экспрессии выявленных новых генов, кодирующих белок.

Выводы: Получены два пантранскриптома кукурузы, содержащие кодирующую (генов кодирующих белок) и некодирующую (днРНК) части. Для обоих пантранскриптомов были выявлены группы ортологов, а также консервативные и вариабельные части. Кроме того, для пантранскриптома, включающего некодирующие регионы, был проведен поиск новых днРНК.

Финансирование: Работа выполнена за счет финансирования Курчатовского геномного центра ФИЦ ИЦиГ СО РАН, соглашение с Министерством образования и науки РФ № 075-15-2019-1662. Вычисления проводились с использованием ресурсов ЦКП «Биоинформатика».

Pan-transcriptome-based long non-coding RNAs prediction and analysis

Pronozin A.^{1, 2*}, Afonnikov D.^{1, 2}

¹ Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

² Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

* pronozinartem95@gmail.com

Key words: lncRNA; pan-transcriptome; maize

Motivation and Aim: Long non-coding RNAs (lncRNA) are a class of linear or circular RNA molecules of 200 nucleotides or more in length that do not encode proteins. To date, more than half a million lncRNA sequences have been identified for various organisms.

However, despite the growing number of studies, the structural and functional characteristics of lncRNAs remain poorly understood. This is due to the many factors that need to be considered when identifying lncRNAs. The size of lncRNAs is similar to that of some protein-coding genes, leading to overprediction errors. Also, studies show that lncRNA sequences undergo rapid evolution, the patterns of which have not yet been investigated. Thus, there is a need to analyze the gene composition in several members of a species at once. For this purpose, the concepts of pan-genome and pan-transcriptome have been proposed.

However, to date, the works devoted to the study of pan-transcriptoms are mainly focused on the identification and study of new protein-coding genes. Whereas, there are not many works investigating lncRNAs on the scale of pan-transcriptoms, especially for plants.

This work aims to expand the knowledge of the maize pan-transcriptome by further analyzing lncRNAs.

Methods and Algorithms: 503 libraries of maize inbred lines obtained in [1] were used in this work. A bioinformatics pipeline was developed for pan-transcriptome assembly and analysis, which includes three processing steps: quality control and transcriptome assembly, separation of transcriptomes into coding and non-coding parts, and pan-transcriptome analysis.

Results: Two maize pan-transcriptomes were obtained. The pan-transcriptome containing the coding part has, 33023 protein-coding genes. The pan-transcriptome containing the non-coding part has, 2215493 lncRNAs. For both pan-transcriptomes, annotation was performed, orthologous groups were determined, and the conserved and variable parts were characterized. Also, a search for novel lncRNAs was performed

separately for the pan-transcriptome with the non-coding part by aligning the obtained transcripts to databases of known lncRNAs and analyzing the expression of the identified novel lncRNAs. For the pan-transcriptome with a coding part, GO terms were analyzed and the expression of the identified new protein-coding genes was analyzed.

Conclusion: Two maize pan-transcriptomes containing a coding (protein coding genes) part and a non-coding (lncRNA) part were obtained. For both pan-transcriptomes, orthologous groups were identified, and both conserved and variable parts were identified. In addition, for the pan-transcriptome that includes non-coding regions, new lncRNAs were searched for.

Funding: This work was funded by the Kurchatov Genomic Center of the Institute of Cytology and Genomics of the Siberian Branch of the Russian Academy of Sciences, agreement with the Ministry of Education and Science of the Russian Federation No. 075-15-2019-1662. The calculations were performed using the resources of the Bioinformatics Center.

Список литературы/References

1. Hirsch C.N. et al. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*. 2014;26(1):121-135