

УДК 640

NANOTRF: ПРОГРАММА ДЛЯ DE NOVO ПОИСКА ВЫСОКОКОПИЙНЫХ ТАНДЕМНЫХ ПОВТОРОВ В ДАННЫХ НАНОПОРОВОГО СЕКВЕНИРОВАНИЯ ДНК РАСТЕНИЙ**Е.И. Колганова, И.В. Киров***ВНИИ сельскохозяйственной биотехнологии», Москва, Россия*

Высококопийные тандемно организованные повторы (ТП) могут занимать значительную часть (до 30 %) генома растений. В большинстве случаев они локализованы в (пери) центромерной и (суб) теломерной областях генома и участвуют в структурной организации генома, делении клеток, а также могут кодировать РНК. Функции, происхождение, эволюция и геномная организация ТП слабо изучены. Одной из главных причин этого является сложность сборки полноразмерных кластеров ТП в геноме из коротких ридов, что приводит к сильному уменьшению размера кластера ТП и искажению реального сиквенса. Новым инструментом в секвенировании генома является нанопоровое секвенирование (ONP), которое позволяет получить достаточно протяжённые риды, длина которых ограничена длиной выделенных молекул ДНК и может достигать нескольких сотен тысяч пар нуклеотидов. Число видов растений, для которых выкладываются в свободный доступ данные ONP, растёт в геометрической прогрессии. Однако, из-за высокого процента ошибок секвенирования использование таких данных для прямого поиска ТП обычными инструментами (например, Tandem Repeat Finder, (Benson, 1999)) невозможно.

В данной работе мы разработали специальный подход, названный nanoTRF, который позволяет найти высококопийные ТП, рассчитать их копийность в геноме и собрать их консенсусные последовательности для последующего анализа (например, с помощью FISH). NanoTRF состоит из нескольких модулей и получает первичные данные по ТП в каждом риде ONP с помощью недавно опубликованной программы, TideHunter (Gao et al., 2019). BLAST – модуль проводит реципрокное сравнение всех найденных тандемных повторов в ONP ридов и генерирует таблицу пар схожих ТП. Далее с помощью Louvain-модуля проводится кластеризация всех повторов на основе Лувенского метода (Louvain Method). Filtering – модуль, предназначенный для выбраковки тандемных повторов, и Consensus – модуль для построения консенсуса на основе массива выровненных последовательностей.

В качестве входных данных для запуска nanoTRF требуются данные полногеномного секвенирования Oxford Nanopore Technologies (ONT) в формате FASTQ или FASTA.

Успешность работы инструмента была доказана на данных нанопорового секвенирования *Deschampsia antarctica*. Используя nanoTRF, нами были идентифицированы более 40 ТП и получены их сиквенсы. Ранее в работах по изучению генома данного вида были обнаружены 14 высококопийных тандемных повторов с применением инструмента TAREAN (Novák et al., 2017, González et al., 2018). Мы сравнили обнаруженные ранее тандемные повторы с теми, которые были собраны nanoTRF. Результаты показывают высокую корреляцию между нашими данными и данными TAREAN по числу копий в геноме индивидуальных повторов. Кроме этого, нам удалось выявить 13 новых ТП, чья геномная организация ещё не изучена.

Таким образом, nanoTRF позволяет точно определять высококопийные ТП в «сырых» данных ONP и позволит в дальнейшем существенно увеличить число новых ТП и расширить наши знания об их организации и эволюции. NanoTRF доступен по ссылке: <https://github.com/Kirovez/nanoTRF>.

Работа выполнена при финансовой поддержке Российского Фонда Фундаментальных Исследований (РФФИ грант № 17-00-00336).

ЛИТЕРАТУРА

1. Benson, G. Tandem repeats finder: a program to analyze DNA sequences / G. Benson // Nucleic Acids Research. – 1999. – Vol. 27 – Issue 2. – pp. 573–580
2. Gao, Y. TideHunter: efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain / Y. Gao, B. Liu, Y. Wang, Y. Xing // Bioinformatics. – 2019. – Vol. 35. – Issue 14. – pp. i200-i207
3. González, M.L. Characterization of some satellite DNA families in *Deschampsia antarctica* (Poaceae) / M.L. González, J.O. Chiappella, J.D. Urdampilleta // Polar Biology. – 2018. – Vol. 41. – Issue 3. – pp. 457–468
4. Novák, P. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads / P. Novák, L.A. Robledillo, A. Koblížková, I. Vrbová, P. Neumann, J. Macas // Nucleic Acids Research. – 2017. – Vol. 45. – Issue 12. – P. e111