Секция «Биоинформатика»

Функциональный анализ множественно картированных прочтений при изучении ДНК-РНК интеркатома

Научный руководитель – Жарикова Анастасия Александровна

Косимов Мухаммадфирдавс Назарович

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия

E-mail: mihamsik00@gmail.com

В геноме эукариот, в частности человека, представлено большое разнообразие повторяющихся элементов, которые могут составлять более половины общей длины всего генома [1]. Присутствие повторов усложняет анализ данных высокопроизводительного секвенирования (NGS), поскольку последовательности прочтений из этих областей могут быть короче самого повтора и, следовательно, могут быть картированы на несколько мест в геноме. Большинство существующих алгоритмов для анализа NGS-данных не способны эффективно обрабатывать такие множественные картировки, что влечет за собой потерю существенной части информации и затрудняет биологическую интерпретацию результатов.

Существующие референсные сборки геномов эукариот практически не включают локусы, содержащие такие типы повторов, как, например, центромерные и теломерные повторы. Проблема определения последовательности повторяющихся элементов была решена с разработкой протоколов секвенирования третьего поколения. Например, геномная сборка человека T2T (telomere to telomere) была получена с использованием этих протоколов и включает в себя полный набор тандемных и центромерных повторов [2]. Тем не менее, многие распространенные в практике протоколы секвенирования дают короткие чтения. Например, протокол Red-C [3] для анализа полногеномного интерактома между ДНК и РНК генерирует прочтения длиной от 80 до 133 нуклеотидов. Стандартные биоинформатические программные конвейеры, используемые для анализа РНК-ДНК интерактомов подразумевают использование только уникально-картированных прочтений, что приводит к потере трех типов контактов: уникальная ДНК-множественная РНК, множественная ДНК-уникальная РНК, множественная ДНК-множественная РНК. В различных экспериментах потеря данных, связанная с игнорированием множественно картированных чтений, может составлять более половины всех контактов.

Несмотря на неоднозначность происхождения множественно картированных прочтений, было принято решение использовать их в анализе данных ДНК-РНК контактов для выявления функциональной значимости повторяющихся элементов. Был проведен анализ данных, полученных с использованием протоколов для установления полногеномного РНК-ДНК интерактома Red-C и GRID-seq. В качестве референсного генома была использована наиболее полная сборка генома человека (T2T). В ходе анализа разработан протокол, позволяющий использовать информацию о множественном картиорвании РНК- и ДНК-частей контактов. Предложенный подход позволил дополнить информацию о хроматин-ассоциированных РНК, закодированных в геноме в нескольких копиях. Показано, что с геномными локусами, несущими повторяющиеся элементы разных классов, взаимодействуют хроматин-ассоциированные РНК, а также удалось выявить тенденции взаимодействия уникальных и множественно картируемых прочтений, приходящихся на различные гены и типы повторов.

[1]Repetitive Elements in Humans

[2]A complete telomere-to-telomere assembly of the maize genome

[3]Studying RNA-DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics