Секция «Биоинформатика»

Автоматизация сборки митохондриальной ДНК содержащей дупликации и тандемные повторы

Научный руководитель – Комиссаров Алексей Сергеевич

Зилов Данил Сергеевич

Студент (магистр)

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия

E-mail: zilov@scamt-itmo.ru

Митохондриальная ДНК (мтДНК) - кольцевая молекула, расположенная в митохондриях эукариот. Особенностью мтДНК является ее незначительное изменение при передачи в поколениях и стабильная структура у организмов одного вида. Эта молекула удобна для определения видовой принадлежности, поэтому ее используют в филогенетике.

В последние несколько лет с развитием технологий секвенирования появилось большое количество данных секвенирования различных животных, для которых мтДНК ранее не была собрана. Мы столкнулись с тем, что для некоторых видов сборка мтДНК является сложной задачей с которой не справляются доступные инструменты. Главной причиной этого является то, что мтДНК этих видов содержат дупликации и тандемные повторы в контрольном регионе. Дополнительно задача сборки мтДНК осложнена тем, что части мтДНК могут встраиваться в геном, образуя так называемые NUMT (nuclear mitochondrial DNA).

Современные инструменты для сборки мтДНК из сырых данных (например NOVOPlasty, Norgal) не справляются с качественной обработкой повторов, что приводит к схлопыванию дуплицированных участков ДНК. Для решения этой проблемы мы решили создать собственный инструмент.

Главным преимуществом предложенного нами подхода является использование индекса построенного на несобранных ридах Illumina основанного на идеальной функции хэширования. Использованный индекс позволяет как получать риды содержащие интересующие нас последовательности мтДНК, так и использовать классическую сборку контигов с использованием графа de Bruijn. Сборка состоит из следующих этапов:

1) фильтрация сырых данных секвенирования на наличие адаптеров и технических последовательностей;
2) удаление оптических дупликатов;
3) построение индекса;
4) поиск затравочных k-меров для инициализации сборки с использованием известных собранных мтДНК;
5) использования затравочных k-меров для сборки унитигов с учетом разницы в покрытие между мтДНК и геномными фрагментами NUMP;
6) разрешение дупликаций повторов с использованием индекса по ридам.

В случае наличия близкого референсного генома возможна сборка мтДНК с использованием графа de Bruijn с учетом референса.

Из-за того, что мтДНК часто представлена в ДНК более обильно, чем геномная, предложенный подход позволяет собирать мтДНК последовательность для из сырых данных с геномным покрытием меньше единицы.

Использованный нами подход был успешно использован для сборки мтДНК в образцах Aratinga sp. (дупликация и тандемный повторы), Psittacus sp. (2 т.н.п. дупликация), черноногих хорьков (короткий однородный тандемный повтор) и ящериц семейства Darevskia sp. (длинный тандемный повтор). Для части образцов сборка была верифицирована длинными ридами PacBio.